

# Computational Approaches to Digitised Historical Newspapers

Maud Ehrmann<sup>\*1</sup>, Marten Düring<sup>\*2</sup>, Clemens Neudecker<sup>\*3</sup>, and Antoine Doucet<sup>\*4</sup>

- 1 EPFL – Lausanne, CH. [maud.ehrmann@epfl.ch](mailto:maud.ehrmann@epfl.ch)
- 2 University of Luxembourg, LU. [marten.during@uni.lu](mailto:marten.during@uni.lu)
- 3 Staatsbibliothek zu Berlin, DE. [clemens.neudecker@sbb.spk-berlin.de](mailto:clemens.neudecker@sbb.spk-berlin.de)
- 4 University of La Rochelle, FR. [antoine.doucet@univ-lr.fr](mailto:antoine.doucet@univ-lr.fr)

---

## Abstract

Historical newspapers are mirrors of past societies, keeping track of the small and great history and reflecting the political, moral, and economic environments in which they were produced. Highly valued as primary sources by historians and humanities scholars, newspaper archives have been massively digitised in libraries, resulting in large collections of machine-readable documents and, over the past half-decade, in numerous academic research initiatives on their automatic processing. The Dagstuhl Seminar 22292 “Computational Approaches to Digitised Historical Newspaper” gathered researchers and practitioners with backgrounds in natural language processing, computer vision, digital history and digital library involved in computational approaches to historical newspapers with the objectives to share experiences, analyse successes and shortcomings, deepen our understanding of the interplay between computational aspects and digital scholarship, and discuss future challenges. This report documents the program and the outcomes of the seminar.

**Seminar** July 17–22, 2022 – <http://www.dagstuhl.de/22292>

**2012 ACM Subject Classification** Computing methodologies → Information extraction; Computing methodologies → Machine learning; Information systems → Digital libraries and archives; Applied computing → Arts and humanities; Applied computing → Document management and text processing; Information systems → Information retrieval; Information systems → Data mining; Information systems → Document representation; Information systems → Document structure; Information systems → Structure and multilingual text search; Information systems → Users and interactive retrieval

**Keywords and phrases** historical document processing, document structure and layout analysis, natural language processing, information extraction, natural language processing, digital history, digital scholarship

**Digital Object Identifier** 10.4230/DagRep.12.7.112

---

\* Editor / Organizer



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Computational Approaches to Digitised Historical Newspapers, *Dagstuhl Reports*, Vol. 12, Issue 7, pp. 112–179  
Editors: Maud Ehrmann, Marten Düring, Clemens Neudecker, and Antoine Doucet



Dagstuhl Reports  
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

## 1 Executive Summary

*Maud Ehrmann (EPFL – Lausanne, CH)*

*Marten Düring (Luxembourg Centre for Contemporary and Digital History, LU)*

*Clemens Neudecker (Staatsbibliothek zu Berlin, DE)*

*Antoine Doucet (University of La Rochelle, FR)*

**License** © Creative Commons BY 4.0 International license  
© Maud Ehrmann, Marten Düring, Clemens Neudecker, and Antoine Doucet

### Context

For long held on library and archive shelving, historical newspapers are undergoing mass digitisation and millions of facsimiles, along with their machine-readable content captured via optical character recognition (OCR), are becoming accessible via a variety of online portals.<sup>1</sup> While this represents a major step forward in terms of preservation and access, it also opens up new opportunities and poses timely challenges for both computer scientists and humanities scholars [2, 1, 3, 14].

As a direct consequence, the last ten years have seen a significant increase of academic research on historical newspaper processing. In addition to decisive grassroots efforts led by libraries to improve OCR technology,<sup>2</sup> individual works dedicated to the development and application of tools to digitised newspaper collections have multiplied [12, 13, 10, 11], as well as events such as evaluation campaigns or hackathons [8, 7, 6, 5].<sup>3</sup> Besides, several large consortia projects proposing to apply computational methods to historical newspapers at scale have recently emerged, including ViralTexts<sup>4</sup>, Oceanic Exchanges<sup>5</sup>, *impresso* – Media Monitoring of the Past<sup>6</sup>, NewsEye<sup>7</sup>, Living with Machines<sup>8</sup>, and DATA-KBR-BE<sup>9</sup> [9].

This momentum can be attributed not only to the long-standing interest of scholars in newspapers coupled with their recent digitisation, but also to the fact that these digital sources concentrate many challenges for computer science, which are all the more difficult – and interesting – since addressing them requires taking digital (humanities) scholarship needs and knowledge into account. Within interdisciplinary frameworks, various and complementary approaches spanning the areas of natural language processing, computer vision, large-scale computing and visualisation, are currently being developed, evaluated and deployed. Overall, these efforts are contributing a pioneering set of tools, system architectures, technical infrastructures and interfaces covering several aspects of historical newspaper processing and exploitation.

<sup>1</sup> Such as those discussed in [16] and by this seminar’s working group on transparency and fairness, see Section 4.3 hereafter.

<sup>2</sup> See e.g., the OCR-D project, an ecosystem for improving OCR on historical documents: <https://ocr-d.de/en/about> and [4].

<sup>3</sup> See the 2017 edition of the Coding Da Vinci cultural hackathon or the 2019 edition of the Helsinki Digital Humanities Hackathon.

<sup>4</sup> A project aiming at mapping networks of reprinting in 19th-century newspapers and magazines (US, 2012-2016): <https://viraltxts.org>

<sup>5</sup> A project tracing global information networks in historical newspaper repositories from 1840 to 1914 (US/EU, 2017-2019): <https://oceanicexchanges.org>

<sup>6</sup> A project which tackles the challenge of enabling critical text mining of newspaper archives (CH, 2017-2020): <https://impresso-project.ch>

<sup>7</sup> A digital investigator for historical newspapers (EU, 2018-2022): <https://www.newseye.eu>

<sup>8</sup> A project which aims at harnessing digitised newspaper archives (UK, 2018-2023): <https://www.turing.ac.uk/research/research-projects/living-machines>

<sup>9</sup> A project which aims at facilitating data-level access to KBR’s collections (mainly newspapers) for open science (2020-2022): <https://www.kbr.be/en/projects/data-kbr-be/>.

## Objectives

The aim of the seminar was to bring together researchers and practitioners involved in computational approaches to historical newspapers to share experiences, analyse successes and shortcomings, deepen our understanding of the interplay between computational aspects and digital scholarship, and begin to design a road map for future challenges. Our seminar was guided by the vision of methodologically reflected, competitive and sustainable technical frameworks capable of providing an extensive, sophisticated and possibly dynamic access to the content of digitised historic newspapers in a way that best serves the needs of digital scholarship. We are convinced that in order to meet the many challenges of newspaper processing and to accommodate the demands of humanities scholars, only a global and interdisciplinary approach that looks beyond technical solutionism and embraces the complexity of the source and its study can really move things forward.

## Participants and Organisation

The seminar gathered 22 researchers<sup>10</sup> with backgrounds in natural language processing, computer vision, digital history and digital library, the vast majority of whom had previously worked on historical newspapers and were familiar with interdisciplinary environments. To structure and coordinate the work of the seminar, the organisers proposed a mixture of plenary sessions, working groups, and talks, as follows (see also Fig.1):

- **Spotlight talks** on day 1, where each participant briefly introduced him/herself and gave an opinion or statement on his/her current view of the main topic of the seminar (3-minute/1-slide).
- **Demo session**, where some participants shortly introduced a relevant asset (e.g., a dataset, tool, interface, on-going experiment).
- **Working group sessions**, during which groups composed of computer scientists and humanities scholars focused on a specific question. Work within a group featured different moments, with: expert group discussion, where people with similar backgrounds exchanged in order to align their understanding of the question at hand and to prioritise problems; observation of concrete research and workflow practices on existing approaches and/or tools; cross-interviews, where people from one domain interviewed one person from another domain about a specific point; mixed group discussion, where everybody jointly reflected; and writing time, where the group wrote a report summarising its findings.
- **Reporting sessions**, where working groups reported their discussion and presented their main conclusions and recommendations in a plenary session.
- **Morning presentations**, where researchers shared their experience from a project and/or their view on a specific topic, followed by a discussion with the participants.
- **Evening talks**, where researchers shared their experience and views on a topic at large. We proposed three evening lectures that addressed the field of digitised newspapers from the perspective of computer science, digital history, and digital libraries.

---

<sup>10</sup>Some participants had to cancel at the last moment due to the pandemic; we thank them for their initial commitment and hope that there will be future opportunities.

	Monday	Tuesday	Wednesday	Thursday	Friday
9:00 - 10:00	- General introduction	Morning Talks (2* 20+10min)	Demo Session	Group Reporting	Group Reporting & Working Group Session
10:00 - 11:00	- Spotlight talks (ca. 22*3min) with coffee break - Topics presentation and expressions of interest	Group Reporting (30')		Working Group Session	
11:00 - 12:00	- Identification of groups <i>(please refer to the Wiki for detailed schedule)</i>	Working Group Session	Working Group Session		
12:15 - 13:30	<i>lunch</i>				<i>end of seminar</i>
13:30 - 15:30	Room assignment and working group session	Working group session	<i>excursion</i>	Working Group Session	
15:30 - 16:00	<i>coffee break</i>	<i>coffee break</i>		<i>coffee break</i>	
16:00 - 17:30	Working Group Session	Working Group Session		Working Group Session	
17:30 - 18:00					
18:00 - 19:00	<i>dinner</i>	<i>dinner</i>		<i>dinner</i>	
20:15 - 21:15	<i>Evening talk (computer science perspective)</i>	<i>Evening talk (digital history perspective)</i>	<i>outside dinner</i>	<i>Evening talk (digital library perspective)</i>	

■ **Figure 1** Schedule of the seminar.

## Topics

The topics and modus operandi of the seminar were not set in stone but discussed and validated with all participants during the first day. First, the organisers proposed three main topics (and several corresponding sub-questions) for the participants to discuss and reflect on during the seminar. On this basis, participants were then invited to express the specific themes, questions and issues they wished to work on, in a traditional post-it session. Finally, these propositions were examined and structured by the organisers, who defined four working groups.

## Proposed topics

As a starting point, organisers proposed to consider three closely intertwined topics, which are detailed below to further illustrate the background knowledge of this seminar.

1. **Document Structure and Text Processing.** While recent work on the semantic enrichment of historical newspapers has opened new doors for their exploration and data-driven analysis from a methodological perspective (e.g., n-grams, culturomics), results up to now often confirmed common knowledge and were not always considered relevant by historians. The next natural and eagerly-awaited step consists in enriching newspaper contents and structure with semantic annotations which allow for the exploration of far more nuanced research questions. In this regard, several issues arise, among others:
  - **Q1.1 – Complex structures and heterogeneity of contents.** Newspapers are typically composed of a diverse mix of content including text, image/graphical elements, as well as tabular data and various other visual features. The proper segmentation of the page content into individual information pieces is key for enabling advanced research and analysis. This includes the modelling and detection of logical units on the document (or specifically, issue) level as e.g. articles can span across multiple pages. Also of high relevance to researchers is the more advanced classification and semantic labelling of content units, separating categories such as information, opinion,



stock market indices, obituaries, humour, etc. Despite a growing interest, a good understanding of these complex structures as well as methods and technologies for identifying, classifying and accessing diverse content types through appropriate data models and search interfaces are still lacking.

- **Q1.2 – *Diachronic processing.*** Historical newspaper material poses severe challenges for computational analysis due to their heterogeneity and evolution over time. At language level, besides historical spelling variation which leads to major problems in text recognition and retrieval, sequential labelling tasks such as named entity recognition and disambiguation are problematic and often require time-specific resources and solutions. At document level, text classification or topic modelling need to pay attention to the necessary historical contextualisation of their category schemes or corpus time-spans in order to avoid anachronisms. Finally, at structure level, layout processing faces similar challenges and its application needs to adapt to changing sources.
2. **Visualisation, Exploitation and Digital Scholarship.** Historians and other user groups require tools for content discovery and management to reflect their iterative, exploratory research practices. The opportunities and challenges posed by mass digitised newspapers and other digitised sources require them to adjust their current workflows and to acquire new skills.
    - **Q2.1 – *Transparency and digital literacy.*** In the context of research, humanists’ trust in computer systems is dependent on sufficient comprehension of the quality of the underlying data and the performance of the tools used to process it. One way to generate such trust is to create transparency, here understood as: information on the provenance and quality of digital sources; information which allows users to make informed decisions about the tools and data they use; and information which allows their peers to retrace their steps. Such transparency empowers users to use the system in a reflective way. But there is to-date no shared understanding of which information exactly is required to achieve transparency: technical confidence scores are themselves hard to interpret and do not translate easily into actionable information. Likewise, historians can not be expected to be aware of the consequences of all the algorithmic treatments to which their digitised sources have been exposed. Instead, the identification of the most relevant biases and their concrete consequences for users appears to be a more realistic approach. Once these are understood, counter-action can be taken.
    - **Q2.2 – *Iterative content discovery and analysis.*** In contrast to many other applications in computer science, the discovery of relevant content is of greater interest to historians than the detection of patterns in datasets following a priori hypotheses. Historical research is typically iterative: The study of documents yields new insights which determine future exploration strategies and allow scholars to reassess the value of the sources they have consulted. Semantically enriched content offers multiple ways to support this iterative exploration process. New tools for content discovery also require “generous” interfaces, i.e. interfaces which allow users to discover content rather than relying on narrow keyword search [15].
  3. **System Architecture and Knowledge Representation.** The application of various natural language processing and computer vision components which transform noisy and unstructured material into structured data also poses challenges in terms of system architecture and knowledge representation. If those two well-studied fields already offer a strong base to build upon, many questions arise from newspaper source specificities and

the digital humanities context.

- **Q3.1 – *Managing provenance, complexity, ambivalence and vagueness.*** Lots of factual and non factual information is extracted from newspaper material and need to be stored and interlinked. In this regard, two points require great attention. First, newspapers – like any other historical source – represent past realities which do not necessarily align with present-day realities: institutions and countries change names or merge, country borders move or become disputed and professions change or disappear. These temporal shifts, ambivalences and contradicting information cause historical data to be highly complex and sometimes disputed, and the representation of this complexity poses interesting challenges for computer science. Second, if processing steps, and possibly intermediary representations, of algorithms are recorded for the purpose of transparency, this meta-knowledge need to be stored alongside the data.
- **Q3.2 – *Dynamic processing.*** Historical newspaper processing outputs are useless if not used by scholars who wish to investigate research questions. If all methods and practices can not be transposed as they stand from analogue to digital, careful consideration must be given to how best to accommodate scholarship requirements in digital environments where primary sources are turned into data.

### Selected Topics and Working Groups

The discussion around topics led to the definition of four working groups which the participants joined on the first day (on a voluntary basis) and in which they worked throughout the week. No guidelines were given and the groups were free to adapt the direction of their work. Each group wrote a report summarising their activities and findings in Section 4.

1. **Working Group on Information Extraction.** Initiated around the topic of information extraction, this group eventually settled on the specific topic of person entity mentions found in historical newspapers but not present in knowledge bases, a.k.a “hidden people”. The group defined a number of challenges and worked – in a productive hackathon style – on several experiments (see Section 4.1).
2. **Working Group on Segmentation and Classification.** Members of this group quickly discarded the segmentation question to focus on classification only, considering classification scope and practices in relation to digitised newspapers (see Section 4.2).
3. **Working Group on Transparency, Critics and Newspapers as Data.** This group (the largest) worked on a set of recommendations regarding the different aspects of transparency and fairness needed for the analysis of digitised and enriched historical newspaper collections (see Section 4.3).
4. **Working Group on Infrastructure and Interoperability.** This group discussed the issue of consolidation, growth and sustainability of current and future achievements in digitisation, access, processing and exchange of historical newspapers (see Section 4.4).

### Spotlight Talks on the Main Challenges Ahead for Digitised Historical Newspapers

On the first morning of the seminar, the organisers asked participants to briefly present their views on some questions they had to consider in advance. These questions were:

- What are the main challenges we need to address in relation with historical newspapers?
- What is the most exciting opportunity you would like to explore during this seminar?

- If you were given €1 million to spend in the next 6 months on historical newspapers, what would you do?

As well as being a good ice-breaker and kick-off to the seminar, the series of responses to these questions documents what a community of researchers in July 2022 believe to be the next challenges for computational approaches to historical newspapers. In total, 21 researchers formulated no less than 67 statements as responses to the first question. In what follows, we provide a summary of the main ideas and suggestions which we have grouped in 8 themes that cover more or less the whole spectrum of activities around digitised newspapers. Apart from this grouping, no further reflection or refactoring has been done on these statements. While most of the answers are not a surprise to those familiar with the subject, they confirm existing needs, reflect on-going trends, and reveal new lines of research.

► **Document processing.** A first group of statements relates to optical character recognition and optical layout recognition (OLR), two critical processes when working with newspapers. These two document image refinement techniques are extremely difficult when applied to such sources (especially for collections digitised long ago), which explains why they are still high on the agenda despite all the efforts invested in recent years. The views expressed highlight and confirm several dimensions, namely: OCR and OLR quality needs to be improved, finer-grained segmentation and classification of news items is necessary, and processes should be more robust across time and collections. Intensive work is being carried out in these areas.

#### Verbatim statements:

- *How to make available digitised newspaper collections in high quality OCR+layout* (Clemens Neudecker);
- *High quality article segmentation and classification* (Maud Ehrmann, Mickaël Coustaty);
- *The (massive) segmentation bottleneck* (Antoine Doucet);
- *Robust layout recognition* (Matteo Romanello);
- *A level playing field: OCR and fine-grained content segmentation* (Marten Düring);
- *Improving article segmentation (e.g., wrt advertisements and classifieds)* (Mariona Coll-Ardanuy);
- *Layout recognition (e.g., article separation, recognition of headings and authors' names)* (Dario Kampkaspar);
- *OCR+, layout recognition, article segmentation and classification* (Eva Pfanzelter);
- *Better article segmentation, and a way to deal with heterogeneous qualities of segmentation in DL* (Axel Jean-Caurant);
- *Quality of OCR across periods, languages and original document qualities* (Yves Maurer);
- *Standardised approaches to segment historical newspaper pages* (Stefan Jänicke).

► **Text and image processing.** This group encompasses all types of content processing applied to OCR and OLR outputs in view of enriching newspaper contents with further information, usually in the form of semantic annotations and item classification. The main challenges that emerge are: robustness (i.e. approaches that perform well on challenging, noisy input), finer-grained information extraction, few-shot learning (to compensate lack of training data), transferability (approaches that perform well work across settings), multilinguality, multimodality, entity linking, interlinking of collections, and transmedia approaches.

#### Verbatim statements:

- *Robust multilingual information extraction* (Maud Ehrmann);
- *Developing methods that are robust to OCR errors* (Mariona Coll-Ardanuy);

- *Words with meaning change over time* (Martin Volk);
- *Text summarisation and text classification (monolingual and across languages)* (Martin Volk);
- *How to automatically detect genres (in particular film reviews and film listings)* (Julia Noordeggraaf);
- *Multilinguality* (Eva Pfanzelter);
- *Ease multilingual scholarship (qualitative and quantitative)* (Antoine Doucet);
- *Investigating the relation between (or intertwining of) image and text* (Kaspar Beelen);
- *Embedding newspaper content within the media landscape* (Kaspar Beelen);
- *Data mining in newspapers (e.g. biographies, TV/radio programmes)* (Marten Düring);
- *Robust entity linking (multilingual historical documents)* (Matteo Romanello);
- *Entity Linking and visualisations over time, space and networks* (Martin Volk);
- *Linking with other data sources (parliamentary protocols, wiki-data, (historic) names-db, other newspaper portals, (historic) place names db, etc.)* (Eva Pfanzelter);
- *Create links between newspaper contents (topics, entities) and knowledge bases* (Simon Clematide);
- *Automate content analysis (discourses, argumentation, events, meaning, topics) to enable historical research* (Eva Pfanzelter);
- *Learn with few samples and human interactions* (Mickaël Coustaty).

► **Digitisation and Content Mining Evaluation.** Here we have grouped together views on the evaluation of technical approaches and tools at large, and the means to implement it. Important points that emerge are: better metrics, more and diverse gold standards, and better contextualisation and understanding of (sources of) errors.

#### Verbatim statements:

- *How to arrive at common methods and metrics for quality of digitised newspapers* (Clemens Neudecker);
- *Sustainably sharing ground truth datasets and training models* (Sally Chambers);
- *Developing a variety of NLP benchmarks for different tasks across different languages and types of publications* (Mariona Coll-Ardanuy);
- *Build a general taxonomy of content items (including ads, service tables, etc.) and prepare well-sampled data sets from a variety of publication places and time periods* (Simon Clematide);
- *Disentangling correlation of errors and missingness with time, place, language, network position, etc.* (David Smith).

► **Exploration of (enriched) newspaper collections and beyond.** One of the opportunities that researchers have been working on in recent years is new ways of exploring newspaper content. This group of statements is part of this context and highlights some of the long-awaited next steps: unified access to newspaper collections, support for data-driven research, and connection to other archives.

#### Verbatim statements:

- *Access to newspaper content across collections/projects/platforms* (Matteo Romanello);
- *Unified access to all collections with advanced exploration capacities* (Maud Ehrmann);
- *A unified framework to REALLY make collections accessible, usable and interoperable* (Antoine Doucet);
- *Access across collections and copyright hurdles* (Marten Düring);
- *Silos* (Jana Keck);

- *Data-driven linking and analysis of multiple types of sources (e.g. radio, TV, parliamentary records) and datasets (e.g. land ownership, migration)* (Marten Düring);
- *User-driven (from novice to expert) image, information and metadata etc. extraction* (Eva Pfanzelter);
- *Offer our users more than search, but what? Topics, recommenders, ...?* (Yves Maurer);
- *Contextual information extracted from the corpus: hints on rubrics, themes, top keyword per month etc.* (Estelle Bunout);
- *Comparative analyses of contents (political targets of publishers), ordering of articles and time-based development of topics* (Stefan Jänicke).

► **Working with data.** In addition to working with enriched sources that can be semantically indexed and thus more easily retrieved and analysed, researchers (especially historians) also express the wish to work directly with raw data – digitised documents, annotations, or both – and be able to build their own corpora.

**Verbatim statements:**

- *How to create useful datasets and corpora from digitised newspapers* (Clemens Neudecker);
- *Availability of digitisation output (images, text) for further use (beyond interfaces)* (Estelle Bunout);
- *Newspapers as Data: how to facilitate dataset / corpus building* (Sally Chambers);
- *Ease the building and sharing of corpora, taking into account the context of creation (queries, quality, etc)* (Axel Jean-Caurant).

► **Collections, source and tool criticism; Documentation; Inclusivity.** The validity of any conclusions drawn in empirical research depends on a solid understanding of the data used for the analysis. Digitised and enriched newspapers contain multiple levels of processing which often vary significantly across titles in terms of processing quality and extent of enrichment. The statements below point to key challenges and opportunities to advance our reflected analysis of digitised newspapers.

**Verbatim statements:**

- *How to support and perform source criticism on digitised newspaper collections* (Julia Noordegraaf);
- *Methodological guidelines for the computational analysis of newspaper content* (Julia Noordegraaf);
- *Describing how biases arise in digitised newspapers collections (“full-stack bias”)* (Kaspar Beelen);
- *Understanding how structured missingness and data quality affect (historical) research* (Kaspar Beelen);
- *Selection criteria guidelines for what is being selected, digitised, accessible and how it is represented, searchable, and available* (Jana Keck);
- *Trustable and/or understandable approaches to meet users’ needs* (Mickaël Coustaty);
- *How do we make collections as well as access mechanisms inclusive?* (Laura Hollink);
- *How do we monitor fairness of computational approaches to historic newspapers?* (Laura Hollink);
- *How well does the collection support different user groups?* (Laura Hollink);
- *How do we make perspectives in the data explicit? (e.g in NL context: words signalling a colonial perspective)* (Laura Hollink);
- *Information on the scope, contents and quality of a collection, e.g., included titles, covered time periods, granularity of items (page vs. article), OCR quality, corpus statistics* (Estelle Bunout);
- *Investigate the role of attributes like font face and style, margins, layout, paper, etc.* (Stefan Jänicke);

- *Book-historical studies of editorship and publishing (costs, layout, format, advertising, syndicates, networks) crossing national and cataloguing (newspaper/magazine) boundaries* (David Smith) ;
- *Investigate the role of attributes like font face and style, margins, layout, paper, etc.* (Stefan Jänicke).

► **Workflows.** The combination of multiple processes, moreover between different actors, requires the design of more standardised and efficient workflows encompassing the many processing steps that have emerged in recent years.

**Verbatim statements:**

- *Advanced digitisation workflows: from digitisation to OCR to article segmentation* (Sally Chambers);
- *Workflows that conflate search, annotation, classification, corpus construction* (David Smith).

► **Legal matters.** Finally, a last set of (unavoidable) challenges concerns legal issues, with questions of copyright clearance and management, and of personal data, whether it be user data handled by platforms, or the right to be forgotten.

**Verbatim statements:**

- *Find sustainable ways to work with copyright-restricted data sets* (Yves Maurer);
- *Access across collections and copyright hurdles* (Marten Düring);
- *Copyrights and proprietary rights, image rights etc.* (Eva Pfanzelter);
- *Legal questions (copyright, personal rights, etc.)* (Dario Kampkaspar).

## Acknowledgements

This seminar was originally planned for September 2020 but was cancelled due to the COVID-19 pandemic and rescheduled 2 years later. We would like to thank the administrative and scientific teams at Dagstuhl for their support and professionalism throughout the (re)organisation of this seminar as well as the staff on site for their valuable every day help and care. We also thank all the participants for accepting our invitation to spend a week exchanging views, examining, questioning, debating (and writing) about computational approaches to historical newspapers.

## References

- 1 Bingham, A. The Digitization of Newspaper Archives: Opportunities and Challenges for Historians. *Twentieth Century British History*.21, 225-231 (2010,6)
- 2 Deacon, D. Yesterday’s Papers and Today’s Technology: Digital Newspaper Archives and “Push Button” Content Analysis. *European Journal Of Communication*. 22, 5-25 (2007)
- 3 Nicholson, B. The Digital Turn. *Media History*. 19, 59-73 (2013,2)
- 4 Neudecker, C., Baierer, K., Federbusch, M., Boenig, M., Würzner, K., Hartmann, V. & Herrmann, E. OCR-D: An End-to-End Open Source OCR Framework for Historical Printed Documents. *Proceedings Of The 3rd International Conference On Digital Access To Textual Cultural Heritage*. pp. 53-58 (2019,5)
- 5 Ehrmann, M., Romanello, M., Najem-Meyer, S., Doucet, A. & Clematide, S. Extended Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents. *Proceedings Of The Working Notes Of CLEF 2022 – Conference And Labs Of The Evaluation Forum*. 3180 (2022) <https://infoscience.epfl.ch/record/295816>

- 6 Ehrmann, M., Romanello, M., Flückiger, A. & Clematide, S. Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers. *Working Notes Of CLEF 2020 – Conference And Labs Of The Evaluation Forum*. 2696 pp. 38 (2020)
- 7 Clausner, C., Antonacopoulos, A., Pletschacher, S., Wilms, L. & Claeysens, S. PRImA, DMAS2019, Competition on Digitised Magazine Article Segmentation (ICDAR 2019). (2019), <https://www.primaresearch.org/DMAS2019/>
- 8 Rigaud, C., Doucet, A., Coustaty, M. & Moreux, J. ICDAR 2019 Competition on Post-OCR Text Correction. *2019 International Conference On Document Analysis And Recognition (ICDAR)*. pp. 1588-1593 (2019)
- 9 Ridge, M., Colavizza, G., Brake, L., Ehrmann, M., Moreux, J. & Prescott, A. The Past, Present and Future of Digital Scholarship with Newspaper Collections. *DH 2019 Book Of Abstracts*. pp. 1-9 (2019), <http://infoscience.epfl.ch/record/271329>
- 10 Kestemont, M., Karsdorp, F. & Düring, M. Mining the Twentieth Century's History from the Time Magazine Corpus. *Proceedings Of The 8th Workshop On Language Technology For Cultural Heritage, Social Sciences, And Humanities (LaTeCH)*. pp. 62-70 (2014)
- 11 Lansdall-Welfare, T., Sudhakar, S., Thompson, J., Lewis, J., Team, F. & Cristianini, N. Content Analysis of 150 Years of British Periodicals. *Proceedings Of The National Academy Of Sciences*. **114**, E457-E465 (2017)
- 12 Moreux, J. Innovative Approaches of Historical Newspapers: Data Mining, Data Visualization, Semantic Enrichment. *IFLA News Media Section, Lexington, August 2016, At Lexington, USA*. (2016,8), <https://hal-bnf.archives-ouvertes.fr/hal-01389455>
- 13 Wevers, M. Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950-1990. *Proc. Of The 1st International Workshop On Computational Approaches To Historical Language Change*. (2019), <https://www.aclweb.org/anthology/W19-4712>
- 14 Bunout, E., Ehrmann, M. & Clavert, F. (editors) Digitised Newspapers – A New Eldorado for Historians? Tools, Methodology, Epistemology, and the Changing Practices of Writing History in the Context of Historical Newspaper Mass Digitisation. De Gruyter (2022, in press), doi:10.1515/9783110729214, <https://www.degruyter.com/document/isbn/9783110729214/html>
- 15 Whitelaw, M. Generous Interfaces for Digital Cultural Collections. *Digital Humanities Quarterly*. **9** (2015), <http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html>
- 16 Ehrmann, M., Bunout, E. & Düring, M. Historical Newspaper User Interfaces: A Review. *Proceedings Of The 85th International Federation Of Library Associations And Institutions (IFLA) General Conference And Assembly*. pp. 24 (2019), <https://infoscience.epfl.ch/record/270246?ln=en>



## 2 Table of Contents

### Executive Summary

*Maud Ehrmann, Marten Düring, Clemens Neudecker, and Antoine Doucet* . . . . . 113

### Overview of Talks

Memoirs of Extraordinary Popular Delusions <i>David A. Smith</i> . . . . .	124
Living with Machines: Exploring Newspapers at Scale <i>Kaspar Beelen and Mariona Coll-Ardanuy</i> . . . . .	124
NLP on Historical Documents: Experience (from impresso), Challenges, Opportunities <i>Simon Clemenide</i> . . . . .	125
Integrating Computational Processing into Historical Research and the Steps Leading to it. <i>Estelle Bunout</i> . . . . .	126
Newspapers as Data: Challenges and Solutions <i>Sally Chambers</i> . . . . .	126

### Working groups

Tracking Discourses on Public and Hidden People in Historical Newspapers <i>Simon Clemenide, Mariona Coll Ardanuy, Yves Maurer</i> . . . . .	127
Current Practices of Iterative Classification Approaches for Digitised Historical Newspaper Collections <i>Mickaël Coustaty, Estelle Bunout, Jana Keck, and David A. Smith</i> . . . . .	138
Fairness and Transparency throughout a Digital Humanities Workflow: Challenges and Recommendations <i>Kaspar Beelen, Sally Chambers, Marten Düring, Laura Hollink, Stefan Jänicke, Axel Jean-Caurant, Julia Noordegraaf, and Eva Pfanzerter</i> . . . . .	144
Towards an International Historical Newspaper Infrastructure <i>Clemens Neudecker, Maud Ehrmann, Dario Kampkaspar, Matteo Romanello, Martin Volk, and Lars Wieneke</i> . . . . .	174

**Participants** . . . . . 179

**Remote Participants** . . . . . 179

### 3 Overview of Talks

#### 3.1 Memoirs of Extraordinary Popular Delusions

*David A. Smith (Northeastern University – Boston, US)*

**License** © Creative Commons BY 4.0 International license  
© David A. Smith

**Joint work of** David A. Smith, Ryan Cordell

Newspapers present an extraordinary window into modern language, history, and culture and a revealing mode of information production. By tracing how texts are exchanged, edited, composed, laid out, and generally reprinted, we can learn about historical communications, political, social, and transportation networks. The sample of newspapers digitised by most projects, however, is not representative of all aspects of the historical population. We can correct for this mismatch using regression on observed features of undigitised papers from catalogues and historical directories. Finally, we can use the structure of text reprinting to correct and retrain transcription and layout analysis models.

#### 3.2 Living with Machines: Exploring Newspapers at Scale

*Kaspar Beelen (The Alan Turing Institute – London, GB)*

*Mariona Coll-Ardanuy (The Alan Turing Institute – London, GB)*

**License** © Creative Commons BY 4.0 International license  
© Kaspar Beelen and Mariona Coll-Ardanuy

**Joint work of** The Living with Machines Project

**URL** <https://livingwithmachines.ac.uk/team-2/>

Living with Machines (LwM) is an interdisciplinary research project focused on the lived experience of Britain’s industrialisation during the long nineteenth century (roughly 1780 to 1918). The project develops approaches made possible by rapid digitisation and computational methods, to analyse and link historical records. In this presentation, we focused on our work on historical newspapers. The digitised press provides an immense amount of varied, fine-grained, and often neglected information. However large and rich, newspapers remain difficult to navigate as existing tools struggle with the particularities of digitised historical data.

News was often about place, its discourse anchored in space. In our talk, we show that historically sensitive methods improve performance for both toponym recognition and resolution. Moreover, news content was embedded in social and historical contexts. We presented the “Environmental Scan” which explores questions of representativeness and bias based on insights derived from contemporaneous reference sources about the press such as newspaper press directories.

#### References

- 1 Ardanuy, M., Beavan, D., Beelen, K., Hosseini, K., Lawrence, J., McDonough, K., Nanni, F., Strien, D. & Wilson, D. A Dataset for Toponym Resolution in Nineteenth-Century English Newspapers. *Journal Of Open Humanities Data*. 8 pp. 3 (2022,1)
- 2 Beelen, K., Lawrence, J., Wilson, D. & Beavan, D. Bias and Representativeness in Digitized Newspaper Collections: Introducing the Environmental Scan. *Digital Scholarship In The Humanities*. <https://doi.org/10.1093/llc/fqac037> (2022,7)

- 3 Hosseini, K., Nanni, F. & Coll Ardanuy, M. DeezyMatch: A Flexible Deep Learning Approach to Fuzzy String Matching. *Proceedings Of The 2020 Conference On Empirical Methods In Natural Language Processing: System Demonstrations*. pp. 62-69 (2020,10)

### 3.3 NLP on Historical Documents: Experience (from impresso), Challenges, Opportunities

Simon Clematide (Universität Zürich, CH)

License © Creative Commons BY 4.0 International license  
© Simon Clematide

URL <https://impresso-project.ch>

Joint work of Impresso project team: Maud Ehrmann, Marten Düring, Simon Clematide, Matteo Romanello, Estelle Bunout, Daniele Guido, Philipp Ströbel, Roman Kalyakin, Lars Wieneke, Andreas Fickers, Martin Volk, Frédéric Kaplan

In this talk we discussed the application of NLP techniques on historical newspapers in light of the *impresso* project experience. In a nutshell, the project ‘*impresso* – Media Monitoring of the Past’ tried to answer the question of how best to accommodate text analysis research tools and their usage by humanities scholars. Using a co-design approach involving text miners, UX/UI designers and historian users, we worked on bringing the content of digitised newspaper silos – often consisting of big messy data – into an interface that allows search, exploration and visualisation of the texts and their semantic enrichment<sup>11</sup>. NLP and text mining techniques involve basic IR keyword indexing, OCR improvements (lately visual transformers such as TrOCR) and language identification. Although not rocket science, it has to be done carefully to keep all downstream processing language-aware. Data-driven word embeddings built on the historical corpora help organise the semantic space and support query expansion. Keyphrase extraction uses NLP and word embeddings to summarise content items in the most concise terms. Topic modelling clusters the documents into topic distributions and serves as a search filter and recommender backbone. Lastly, named entities are indexed and linked to Wikidata. Their distribution in the corpus is highly informative for historians. Being able to represent linguistic elements from words to terms, sentences, documents as comparable vectors in multilingually aligned vector spaces will enable semantic search in the future.


#### References

- 1 Ehrmann, M., Romanello, M., Clematide, S., Ströbel, P. & Barman, R. Language Resources for Historical Newspapers: The Impresso Collection. *Proceedings of The 12th Language Resources And Evaluation Conference*. pp. 958-968 (2020,5), <https://www.aclweb.org/anthology/2020.lrec-1.121>
- 2 Romanello, M., Ehrmann, M., Clematide, S. & Guido, D. The Impresso System Architecture in a Nutshell. *EuropeanaTech Insights*. (2020), <https://pro.europeana.eu/page/issue-16-newspapers#the-impresso-system-architecture-in-a-shell>, <https://infoscience.epfl.ch/record/283595>
- 3 Ehrmann, M., Düring, M., Clematide, S., Romanello, M., Bunout, E., Guido, D., Ströbel, P., Kalyakin, R., Wieneke, L., Fickers, A., Volk, M. & Kaplan, F. *Impresso: Historical Newspapers Beyond Keyword Search*. (*in preparation*).

<sup>11</sup> <https://impresso-project.ch/app/>

### 3.4 Integrating Computational Processing into Historical Research and the Steps Leading to it.

*Estelle Bunout (Luxembourg Centre for Contemporary and Digital History, LU)*

License  Creative Commons BY 4.0 International license  
© Estelle Bunout

The efforts in digitising newspapers have been massive in the past decades and many experiences have been gathered by humanists and libraries using and publishing these collections. While the diversity of formats in which their content is accessible remains high, engaged users can collect different and complementary information from different sources. Humanists and in particular historians, need access to the source material but also to contextual information, often provided in form of metadata. The issue relies most commonly not so much in the lack of experience in producing relevant metadata – being derived from catalogue information or content-based computed metadata, but rather in lack of wide habit of using them by researchers and the tendency for some untypical metadata to be incomplete or unstable. Nevertheless, once access and basic contextual information has been made available, many options of operationalising humanities research questions with the support of computational tools, most commonly text mining, open up. In the talk, a case was presented of using a naive Bayes classifier to look for similar texts to anti-modern articles published in Swiss interwar press. This study highlights the need for contextualisation of findings to interpret them properly.

#### References

- 1 Bunout, E. & Düring, M. Collections of Digitised Newspapers as Historical Sources – Parthenos Training. (2019), <https://training.parthenos-project.eu/sample-page/digital-humanities-research-questions-and-methods/collections-of-digital-newspapers-as-historical-sources/>

### 3.5 Newspapers as Data: Challenges and Solutions

*Sally Chambers (Ghent University, BE & KBR, Royal Library of Belgium, Brussels, BE)*

License  Creative Commons BY 4.0 International license  
© Sally Chambers

Digital cultural heritage collections in libraries, archives, and museums are increasingly being used for digital humanities research. However, traditional ways of providing access to such collections, for example through digital library interfaces, are less than ideal for researchers who are looking to build datasets around specific research questions. Originating in the United States, the “Collections as Data” movement was established to encourage cultural heritage professionals to start thinking differently about how they provide access to their collections to facilitate analysis using digital tools and methods. “Collections as Data” encourages the provision of “data-level access” to the underlying files of digitised and born-digital cultural heritage resources to facilitate data analysis by means of tools and methods developed in the field of digital humanities. This presentation explores whether the application of “Collections as Data” to digitised historical newspapers could help facilitate corpus building for digital humanities research.

## References

- 1 Padilla, T., Allen, L., Frost, H., Potvin, S., Roke, E. & Varner, S. Final Report – Always Already Computational: Collections as Data. *Zenodo. Texas Digital Library*. **10** (2019), <https://zenodo.org/record/3152935>
- 2 Chambers, S., Lemmers, F., Pham, T., Birkholz, J., Ducatteeuw, V., Jacquet, A., Dillen, W., Ali, D., Milleville, K. & Verstockt, S. Collections as Data : Interdisciplinary Experiments with KBR’s Digitised Historical Newspapers : A Belgian Case Study. *DH Benelux 2021, Abstracts*. (2021), <http://hdl.handle.net/1854/LU-8712404>
- 3 Tasovac, T., Chambers, S. & Tóth-Czifra, E. Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper. (2020,10), <https://hal.archives-ouvertes.fr/hal-02961317>

## 4 Working groups

### 4.1 Tracking Discourses on Public and Hidden People in Historical Newspapers

*Simon Clematide (Department of Computational Linguistics, Switzerland, [siclemat@cl.uzh.ch](mailto:siclemat@cl.uzh.ch))*

*Mariona Coll-Ardanuy (The Alan Turing Institute, UK, [mcollardanuy@turing.ac.uk](mailto:mcollardanuy@turing.ac.uk))*

*Yves Maurer (National Library of Luxembourg, Luxembourg, [yves.maurer@bnl.etat.lu](mailto:yves.maurer@bnl.etat.lu))*

License  Creative Commons BY 4.0 International license  
© Simon Clematide, Mariona Coll Ardanuy, Yves Maurer

This working group focused on information extraction on historical newspapers. Following an initial brainstorming session with the whole seminar group, we decided to narrow down the topic of information extraction to the more specific question of the coverage of public (notable) and hidden (non-notable) people in newspapers.<sup>12</sup> From the historian’s perspective, it is difficult to identify and study groups of people who are not already notable for something, and work has been undertaken to develop approaches to represent the under-represented [2]. During the seminar, we built on the idea (and hope) that newspapers could serve as a source of stories about and insights into the lives of people who are not in the public eye. Given the group’s interest in working with real datasets and tools, we agreed to conduct actual experiments in the form of a case study to understand if our questions could be addressed in a practical way.

#### 4.1.1 Discussed Problems

Our overarching research question is whether computational methods can help shed some light on how newspapers report on *public* versus *hidden* people. Our aim was to explore ways to address this question computationally. To do so, we broke it into the following methodological questions:

---

<sup>12</sup>We decided to use the terms ‘public’ and ‘hidden’ in this report. The term ‘notable’ is also often used for public figures, especially in the context of Wikipedia. We chose ‘hidden’ not for the reason that they are not mentioned at all in newspapers, but that they cannot be found in (cultural) knowledge bases for famous or public people.

- Can we use the Wikidata linkability of historical newspaper ground-truth datasets on entity linking<sup>13</sup> as a proxy for the distinction of public vs hidden people?
- What type of public figures are actually linked in historical newspaper datasets?
- Are there linguistic cues in the context of person entity mentions in newspapers that are reliable enough to be used to detect whether a name is that of a public or hidden person?
- Can we approximate the Wikidata linkability of person mentions by these contextual linguistic features? In other words, can we avoid linking mentions that are unlikely to be in Wikidata?
- How much training material is necessary to achieve a reasonable performance to classify name mentions into public vs hidden? Can these methods be used to effectively collect mentions of hidden people?
- Where do newspapers typically mention hidden people by name (i.e. proper names, as opposed to definite descriptions)? Where do they mention public figures?
- Is there a way to visualise the occurrences of these mentions with respect to page numbers, layout regions, or just within the content of a page?
- Can we identify and cluster typical content items in large newspaper datasets that mention either mostly public, mostly hidden, or a mixture of public and hidden people?

The main goal of the week was to create a proof-of-concept for a subset of the raised problems. We focused our efforts on prototyping a system that would be able to determine, from the context, whether a person mentioned in a newspaper article is a public figure (i.e. someone present in Wikidata), or a hidden person (someone not represented in Wikidata). Additionally, we decided to experiment with innovative visualisation ideas to better understand and explore where public and hidden figures are mentioned in newspapers.<sup>14</sup>

#### 4.1.2 Related work

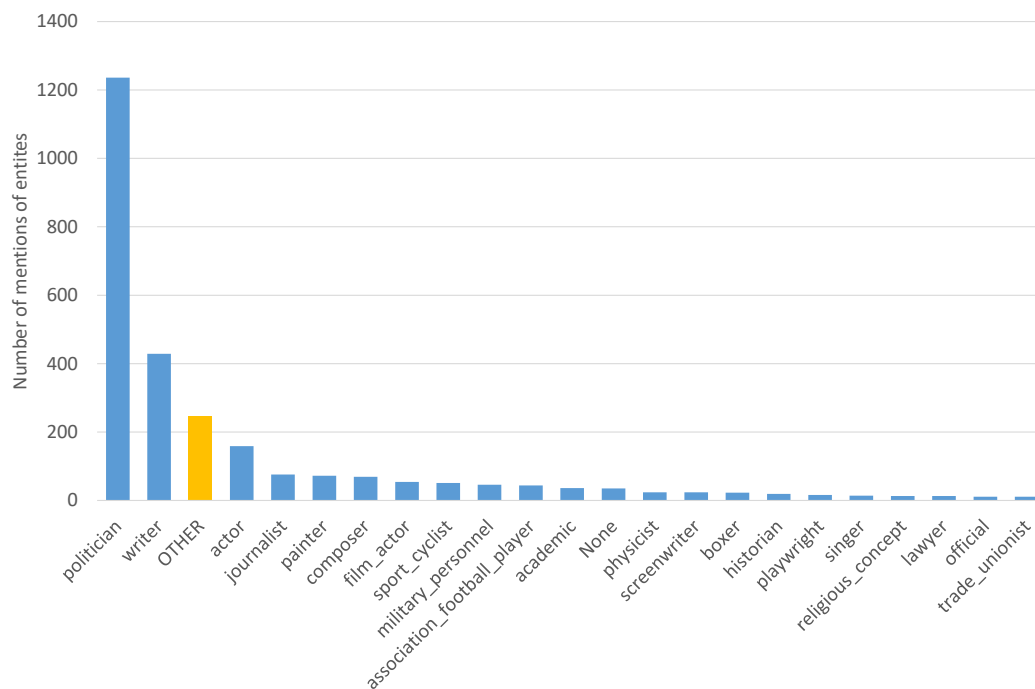
We approached the task of distinguishing between *hidden* and *public* person entities as a sequence tagging problem, basically an extension of the more general named entity recognition and classification task (NERC), which *detects* potential mentions in running text and *classifies* them into different named entity types (such as person, location, and organisation). Ehrmann *et al.* [3] recently surveyed approaches to NERC on historical data.

The task of assigning a unique identifier from a knowledge base (or NIL if the mentioned entity is not represented in the knowledge base) to a mention of a name in running text is known as (named) entity linking (EL), named entity disambiguation/normalisation or, if the link is to Wikipedia or Wikidata, as wikification. A recent survey introduces the approaches to solve this task on modern textual material [4], and the HIPE-2022 overview reports results of a shared task on entity recognition and linking in historical newspaper data in English, German, French, Finnish and Swedish [1].

The topic of notable people in knowledge bases such as Wikidata is presented in [5]. The authors aggregate, validate and analyse the content of different editions of Wikipedia and Wikidata, ending up with a validated subset of 2.29 million persons, from which they conclude

<sup>13</sup>In other words, whether a person mentioned in a news article is present in Wikidata (represented in entity linking datasets with a QID, i.e. a unique Wikidata identifier) or not (represented with ‘NIL’).

<sup>14</sup>We are grateful for the suggestions, ideas and discussions on visualisations that Stefan Jänicke brought into our working group.



■ **Figure 2** Distribution of Wikidata linked entities’ labels (most often their professions in the case of persons) in our French “hipe2020” and “newseye” newspaper dataset. For each person, the chosen profession (among the ones that can be found for the person) is the most common label in the collection (if we take all possible labels for all linked persons into consideration). The OTHER category accumulates entity professions that are mentioned less than 10 times in the dataset.

that it roughly covers an elite of 1/43,000 of humans having ever lived. The distribution of the main occupations of the individuals in this subset is: Culture (31%) including journalists; Sports (28%); Leadership (27%) in politics, religion, nobility etc.; Science/Discovery (12%).

### 4.1.3 Data

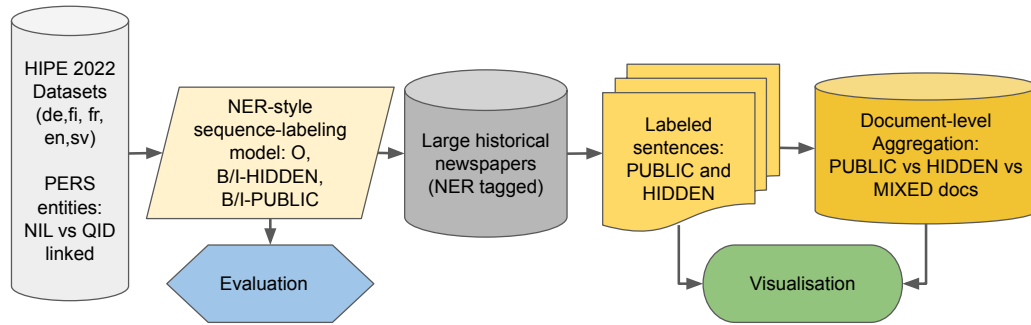
For our experiments, we used two of the French datasets provided as part of the HIPE-2022 shared task on named entity processing in multilingual historical documents<sup>15</sup> [1]: namely the `hipe2020` and `newseye` datasets. They consist of historical newspaper pages or articles in French, ranging from mid-19th century to mid-20th century, annotated for the tasks of named entity recognition and entity linking to Wikidata. In these articles, each entity mention has been manually annotated and classified into one of several categories (including ‘person’), and linked to its corresponding Wikidata QID (i.e. the Wikidata identifier) if existing, or tagged as NIL otherwise.

To better understand the professions of people mentioned in historical newspapers that are linked to Wikidata, and compare it to the distribution of occupations in the Wikidata people dataset mentioned above [5], we analysed the linked person entities mentioned in the `hipe2020` and `newseye` datasets used for this case study.<sup>16</sup> Figure 2 shows the distribution

<sup>15</sup>For more information on the shared task and on the datasets, see <https://hipe-eval.github.io/HIPE-2022/> and <https://github.com/hipe-eval/HIPE-2022-data> respectively.

<sup>16</sup>We used the CLOCQ API [6] for efficient retrieval of the persons’ Wikidata labels.





■ **Figure 3** Proposed experimental pipeline of our case study. Using HIPE-2022 datasets for training a sequence tagger for the recognition of hidden/public persons from NER-tagged input. The results of the application of this model can be visually explored, aggregated on the sentence or document level, and further analysed on the respective levels.

of profession labels of linked persons, with *politician* as the dominant one. Many of the long tail of rare professions are accumulated in the category OTHER. Here it is important to know that a person can have several labels (professions) in Wikidata. For our statistics, we chose for each person the profession that was the most frequent one in our data set. This might explain the imbalanced distribution to some degree.

#### 4.1.4 Approach

Figure 3 shows an overview of the proposed approach and the experiments that we tried to implement. Starting with an entity-linked dataset, we create training and evaluation material by labelling all person entities that are linked to Wikidata as PUBLIC and all entities that are not linked as HIDDEN. In this step, we only consider the proper name parts of the annotations and ignore titles, professions or further specifications attached to the names. The proper name tokens are then masked by generic [PERS] tokens.<sup>17</sup> Next, a traditional NER-style sequence labelling model is trained on the IOB-formatted representation of the training material. The performance of this model on the test set serves as an indicator of whether there is sufficient signal in the linguistic contexts for a reliable distinction between public and hidden. This model can then be applied to further newspaper data. This data needs to be NE-tagged according to the training material, and proper name parts need to be masked by [PERS]. In our case, we just trained an NER tagger with the unmasked training material, but any other tagger could serve the purpose. Finally, given a single newspaper page or any aggregated page subset (e.g. all front pages), we can then visualise where the occurrences of PUBLIC or HIDDEN names are.<sup>18</sup> Another use of the output of the PUBLIC/HIDDEN tagger could be the collection on the sentence level (without layout information playing a role) or on the content item level (e.g., international news vs local news, advertisements, obituaries, radio programs).

<sup>17</sup>We decided to mask person names for two reasons. On the one hand, this was done to force the system to learn from linguistic context, and to avoid the system learning cues that may be implicit or explicit in the person's name. On the other hand, we were aware that public figures appear recurrently in entity linking datasets, and we wanted to avoid our system from basing its predictions on seen dataset-specific training examples.

<sup>18</sup>Alignment between person mention offsets and OCR token image coordinates is required for this.

■ **Table 1** Descriptive statistics of our French datasets compiled from HIPE-2022 datasets `hipe2020` and `newseye`.

Dataset	PUBLIC	%	HIDDEN	%	TOTAL
<code>hipe2020</code>	2143	55.1%	1743	44.9%	3886
<code>newseye</code>	2695	43.7%	3475	56.3%	6170
ALL	4838	48.1%	5218	51.9%	10056

#### 4.1.5 Experiment 1: Hidden vs Public

For the task of HIDDEN vs PUBLIC classification of person names, we fine-tuned a historical BERT model for French (`bert-base-french-europeana-cased`<sup>19</sup>) and worked with HuggingFace’s token classification approach.<sup>20</sup> We fine-tuned the model to the HIPE-2022 data, focusing on the `hipe2020` and the `newseye` datasets. In both datasets, mentions of persons in the data are ideally annotated with a link to Wikidata if the person exists there, and alternatively tagged NIL if the person is absent from Wikidata.<sup>21</sup> Table 1 shows the amount of HIDDEN and PUBLIC entities in our material. Note that overall we have slightly more HIDDEN entities. Interestingly, the two sources have complementary distributions of the two classes: `hipe2020` has more PUBLIC, whereas `newseye` has more HIDDEN. An explanation for this could be that `newseye` pages (full pages were annotated) were randomly sampled, whereas in `hipe2020` the randomly sampled content items (OLR-segmented journal articles) were manually checked before annotation to exclude unsuitable material, such as advertisements. It is also important to note that some of the instances of NIL annotations in the training material are due to the name being ambiguous, or context or time for humans to investigate was insufficient. To balance the proportion between non-person tokens and person tokens, we excluded all sentences in the training material that did not contain any person names. As already mentioned, to facilitate that the model learns from the context, we masked the person names using the [PERS] special token for each proper name token.<sup>22</sup> We used 4,283 sentences for training, 535 for validation, and 536 for evaluation.

#### Findings

We report the results of a single trained model in Table 2. We observe that the performance of this binary classification task in terms of F1 is equal for both classes. However, there is a clear tendency for the model to over-predict PUBLIC compared to HIDDEN. The performance is not perfect overall, but the model clearly finds signal in the context to make the correct prediction in almost 70% of the cases.

<sup>19</sup> <https://huggingface.co/dbmdz/bert-base-french-europeana-cased>

<sup>20</sup> We used the HuggingFace implementation with a learning rate of  $2e - 5$ , a batch size of 16, and weight decay of 0.01, for 10 epochs. We adapted the implementation from the example notebook provided in their Github repository: [https://github.com/huggingface/notebooks/blob/main/examples/token\\_classification.ipynb](https://github.com/huggingface/notebooks/blob/main/examples/token_classification.ipynb)

<sup>21</sup> Note that we take the presence or absence of a person on Wikidata as a proxy of the notability of that person. However, we are aware that this is just a proxy, and we do not consider it a perfect ground truth. A proper evaluation would require annotating the data for the specific objective of distinguishing public and hidden figures.

<sup>22</sup> For example, ‘*M. Pierre Dupond*’ would be masked as ‘*M. [PERS] [PERS]*’.

■ **Table 2** Performance of the PUBLIC-vs-HIDDEN sequential tagger on our test set split for French.

	Precision	Recall	F1
HIDDEN	0.757	0.625	0.685
PUBLIC	0.634	0.744	0.685
Overall	0.688	0.682	0.685

### Examples

Below, we provide some examples of sentences from the test set, in which all proper name parts of person mentions have previously been masked with the special token [PERS]. This is the input to our HIDDEN-vs-PUBLIC sequence labelling model and represents real examples extracted from historical newspaper articles.

► **Example 1.** « [PERS], de la Côte-aux-Fées, fusilier dans la compagnie Germann, et [PERS], fusilier, natif de Neuchâtel ; sont prévenus que s'ils estiment pouvoir prétendre une part aux susdits dons, ils doivent adresser le plutôt possible à la Chancellerie soussignée, les certificats qui constatent les blessures qu'ils ont reçues et les droits qu'ils ont de participer à ces dons, afin que ces pièces puissent être envoyées avant le 15 juillet au comité prémentionné siégeant à Berne. »

→ both mentions are identified as HIDDEN.

► **Example 2.** « Le roi, la reine, le prince héritier, la princesse Louise, le prince Waldemar et sa femme se sont rendus d'hui à bord du vapeur Tantallon-Castle, qui est ancré dans le port de Copenhague, pour rendre visite à [PERS]. »

→ identified as PUBLIC.

► **Example 3.** « M. [PERS] [PERS], mineur, habitant Garnich, a été happé par un convoi de wagonnets Rodange. »

→ identified as HIDDEN.

► **Example 4.** « [PERS], ministre d'Etat belge et chef de la vieille-droite, dont nous avons annoncé le décès à l'âge de 86 ans, était une personnalité de premier plan. »

→ identified as PUBLIC.

► **Example 5.** « [PERS] [PERS], 37 ans, a été poignardé alors qu'il se rendait à l'opéra. »

→ identified as HIDDEN.

► **Example 6.** « Monsieur et Madame [PERS] [PERS] et leur fils [PERS] »

→ all identified as HIDDEN.

► **Example 7.** « Le télégraphe nous apportait lundi matin czar, Alexandre III, a été clamé immédiatement ; c'est le second fils d [PERS] [PERS] »

→ all identified as PUBLIC.

► **Example 8.** « Le Président [PERS] [PERS] a assisté aux funérailles de M. [PERS] [PERS], mineur, habitant Garnich, a été happé par un convoi de wagonnets Rodange. »

→ where the president is identified as PUBLIC, and the miner as HIDDEN.

■ **Table 3** Performance of PERS named-entity tagger trained on the HIPE-2022 French datasets `hipe2020` and `newseye`.

	Precision	Recall	F1
PERS	0.801	0.813	0.811

### Limitations

Whereas in Example 8, the system adequately predicts the president as a public figure and the miner as a hidden figure, based on observation, we find this is a case in which the system struggles, as it appears to push the prediction of one entity in the direction of the other entities in the sentence. Therefore, further quantitative investigation regarding the performance of the tagger in mixed PUBLIC/HIDDEN sentences is needed.

#### 4.1.6 Experiment 2: Applying the Recognition of Public vs Hidden Person Mentions on Other Newspaper Data

In order to test on other newspaper material than HIPE-2022 data, we create a pipeline that first recognises person (PERS) mentions in texts and then classifies them as PUBLIC or HIDDEN. The output of this pipeline serves as input for visualisation experiments. We choose the French-language *Indépendance Luxembourgeoise* newspaper for this experiment because its ALTO XML format contains the page position for each token of the texts. This allows us to explore visualisation ideas that combine HIDDEN/PUBLIC information with layout information.

#### An Application-Specific PERS Tagger

Using the same French HIPE-2022 data as for the HIDDEN/PUBLIC classifier, we trained a HuggingFace model with the same setup to classify PERS entities as HIDDEN or PUBLIC in the Luxembourgish title. The performance of this simple model (F1 score 81%) as shown in Table 3 does not match the state of the art, but we deem it good enough to serve as a NER component in our case study.

### Dataset Description

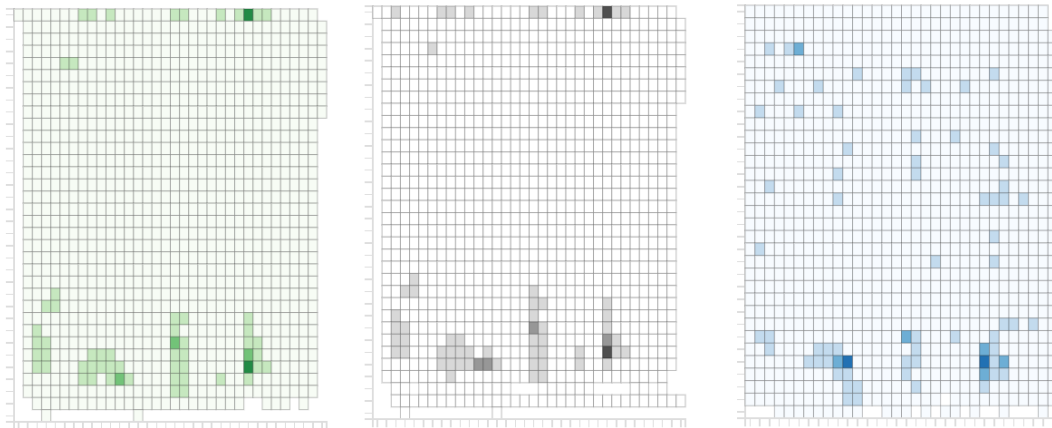
The data used for these experiments is the French-language *Indépendance Luxembourgeoise*<sup>23</sup>. It includes the coordinates of individual words in the ALTO<sup>24</sup> XML format, so the distribution of hidden vs public person entities over pages can be analysed. To examine diachronic changes, two distinct years were selected: 1872 and 1928. See Table 4 for detailed data statistics.

<sup>23</sup><https://eluxemburgensia.lu/periodicals/indeplux>, digitised by the National library of Luxembourg

<sup>24</sup><https://www.loc.gov/standards/alto/>

■ **Table 4** Data statistics and results on *Indépendance Luxembourgeoise*.

Year	Issues	Pages	Sentences	Entities	PUBLIC	HIDDEN
1872	152	608	130,562	17,243	6,166	11,077
1928	141	564	133,512	19,180	7,925	11,166



■ **Figure 4** Distribution of entities over all pages from 1872 and 1928. The green heat map shows all person mentions, the grey heat map renders only HIDDEN ones and the blue heat map the PUBLIC ones.

#### 4.1.7 Visualisations

##### Our Processing and Visualisation Pipeline

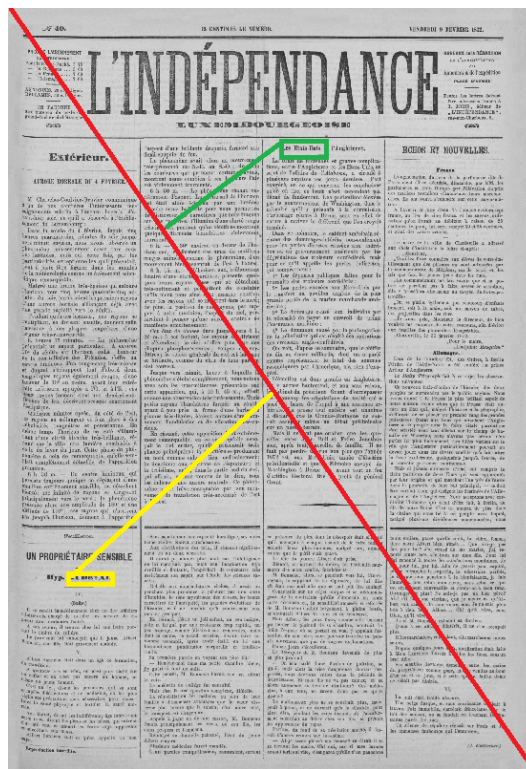
In order to make these visualisations, the ALTO files have been run through the code from `dagstuhl-vips`.<sup>25</sup> The ALTO blocks are first transformed into a JSONL file with the script `extract-alto-blocks.py`. Then the blocks are split into sentences using `spaCy`<sup>26</sup> with `splittextblocks.py`. These sentences are run through our sequence tagger for PERS recognition and HIDDEN/PUBLIC classification. Finally, the tagged sentences from our processing pipeline are combined with the ALTO block data to extract positions of all person mention tokens of HIDDEN/PUBLIC entities on the page by the script `tagged-to-wordpos.py`. Multi-token person names can span more than one line in a newspaper column layout. To keep it simple, we represent each person name by the coordinates of its start token that is tagged either as B-HIDDEN or B-PUBLIC and we refer to them as HIDDEN and PUBLIC respectively. These positions are then loaded into an `elasticsearch`<sup>27</sup> index and visualised using `Kibana`.<sup>28</sup> The visualisations below are all produced from Kibana.

<sup>25</sup> <https://github.com/ymaurer/dagstuhl-vips>

<sup>26</sup> <https://spacy.io/>

<sup>27</sup> <https://www.elastic.co/>

<sup>28</sup> <https://www.elastic.co/kibana/>



■ **Figure 5** Projection of word coordinates onto the top-left to bottom-right diagonal. The top-left position of the token is taken as for person names that span multiple ALTO strings, only the first one is considered. The resulting value is between 0 (top-left) and 1 (bottom right). E.g. the words surrounded by the green and yellow boxes are projected along the red diagonal along the coloured (green and yellow) lines.

**Page Heat Maps**

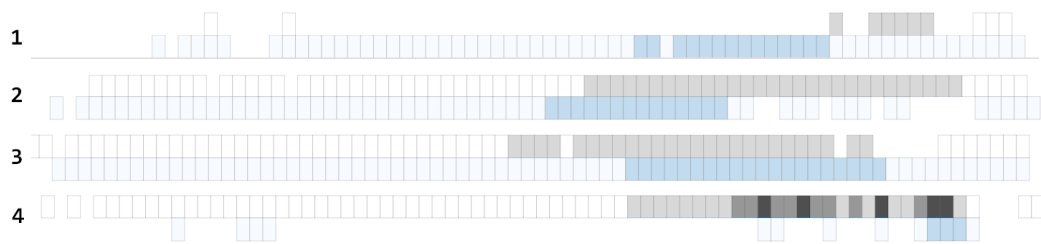
The page heat maps in Figure 4 show that the persons classified as HIDDEN tend to be at the very top or the bottom of the page. The PUBLIC persons are distributed more evenly.

**Projection on the Page Diagonal**

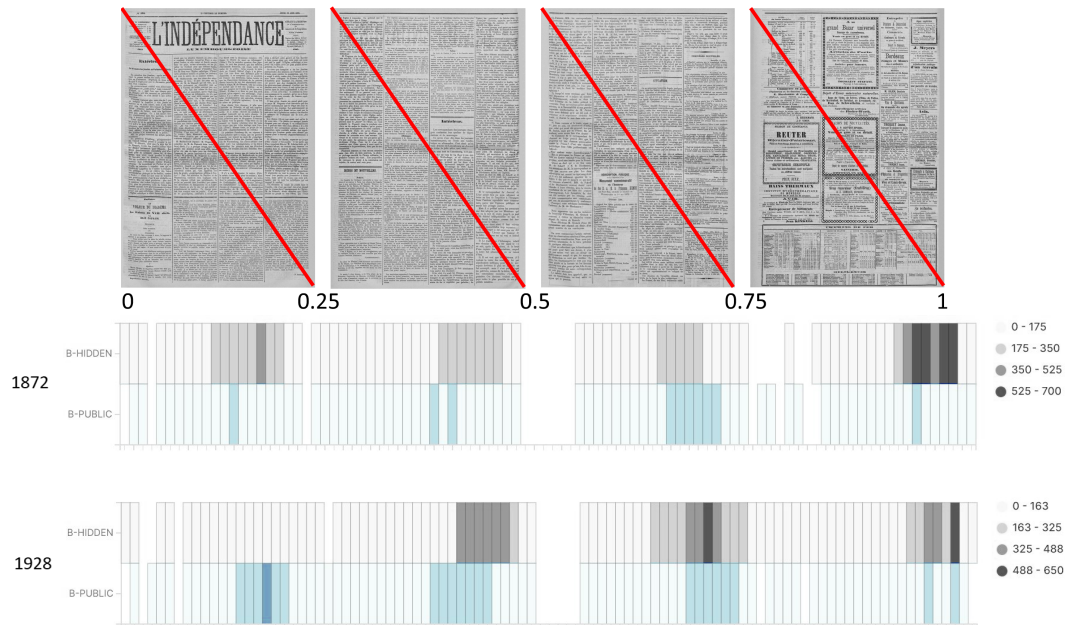
A suggestion by Stefan Jänicke<sup>29</sup> was to simplify the visualisation by projecting the 2-dimensional coordinates of a word onto the 1-dimensional page diagonal. This dimensionality reduction facilitates the visualisation and allows for more abstract comparisons between data points.

The natural reading order for French historical newspapers is from the top-left to the bottom right. The most important information is usually on the top-left because that is the first thing the editor wants the reader to see. Therefore, this projection compresses the dimensions but preserves the order of importance of words to some degree. In that way, all pages can be shown on the same graph or an individual graph per page can be computed as shown in Figure 6. This is simplified by the fact that most newspaper issues at the time had exactly 4 pages.

<sup>29</sup> Associate Professor at the Department of Mathematics and Computer Science at the University of Southern Denmark and participant to this seminar <https://imada.sdu.dk/~stjaenicke/>



■ **Figure 6** Heat map of HIDDEN (grey) and PUBLIC (blue) entities aggregated per page number.



■ **Figure 7** Distribution of entities over pages 1 to 4 from 1872 and 1928. The horizontal axis spans from the top left of page 1 to the bottom right of page 4 as illustrated in the upper part of the figure. The top heat map shows 1872 and the bottom one 1928, the grey part the entities of type HIDDEN and the blue part the entities of type PUBLIC.

### Per Page Projection

Figure 6 illustrates that the persons and their PUBLIC/HIDDEN categories are not distributed uniformly on the pages, and page 4 is notably different from the others. There are several possible interpretations to this. First, the last page contains plenty of advertisements for local companies, classifieds and civil registries; these result in a large number of identified HIDDEN entities. Secondly, on the first page, there are clearly more PUBLIC entities, and they tend to be closer to the top left. We assume that the editors lead with stories about public figures because those interest readers more.

### Diachronic Comparison

The projection along the diagonal can be generalised to a projection over the diagonals of all pages. As shown in Figure 7, this results in a single heat map where the first quarter represents page 1, the second one page 2, etc. This allows a quick visual comparison between different subsets of the data and is therefore particularly suitable for comparing data distributions



for different years (i.e., one can select any set of entity positions on newspaper pages and compare it to another set). The heat maps of Figure 7 show that the distribution of named entities of type person is different for the two years examined, and the distribution of PUBLIC and HIDDEN are also different.

In particular, HIDDEN entities are clustered on the bottom of page 4 for the year 1872, but in 1928 they are on both page 3 and page 4. There are in general more person entities on page 3 in 1928. This again is probably explained by advertisements, which have become more numerous in 1928 and are also filling up the bottom of page 3.

Another notable difference is that on the first pages of 1928 there are more PUBLIC entities than for 1872. There is no clear explanation why this should be the case.

A further difference of a smaller magnitude is the fact that the frequencies of entity types on page 1 are inverted between 1872 and 1928. While 1872 has more hidden entities and few public ones, it's nearly the opposite for 1928. This could be the result from different editorial choices by the editors, by a different layout or by the use of different language around public and hidden figures 56 years apart.

#### 4.1.8 Conclusion and Open Problems

Our case study on French material indicates that linguistic contexts can help identifying person mentions as either public or hidden. Additionally, we show that innovative visualisation techniques could shed light on the distribution of these two categories of peoples in terms of layout position and page number. As mentioned throughout this report, this is an exploratory case study (conducted in a few days during the seminar), and much more needs to be done to reproduce or improve these findings on further material and languages, such as creating task-specific annotations for the distinction between hidden and public. Another point to investigate are ensemble approaches that can help to deal with the model variance that we observed during training and to improve the performance of the public/hidden classification generally.

#### References

- 1 Ehrmann, M., Romanello, M., Najem-Meyer, S., Doucet, A. & Clemenide, S. Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents. *Experimental IR Meets Multilinguality, Multimodality, And Interaction. Proceedings Of The Thirteenth International Conference Of The CLEF Association (CLEF 2022)*. (2022)
- 2 Stranisci M.A., Paii, V. & Damiano R. Representing the under-represented: A dataset of post-colonial, and migrant writers. *3rd Conference on Language, Data and Knowledge, LDK 2021, OpenAccess Series in Informatics*, vol 93, pp 1-14 (2021), <https://dx.doi.org/10.4230/OASICS.LDK.2021.7>
- 3 Ehrmann, M., Hamdi, A., Pontes, E., Romanello, M. & Doucet, A. Named Entity Recognition and Classification on Historical Documents: A Survey. (arXiv,2021), <https://arxiv.org/abs/2109.11406>
- 4 Shen, W., Li, Y., Liu, Y., Han, J., Wang, J. & Yuan, X. Entity Linking Meets Deep Learning: Techniques and Solutions. *IEEE Transactions On Knowledge And Data Engineering*. pp. 1-1 (2021)
- 5 Laouenan, M., Bhargava, P., Eyméoud, J., Gergaud, O., Plique, G. & Wasmer, E. A cross-verified database of notable people, 3500BC-2018AD. *Scientific Data*. **9**, 290 (2022,6,9), <https://doi.org/10.1038/s41597-022-01369-4>
- 6 Christmann, P., Saha Roy, R. & Weikum, G. Beyond NED: Fast and Effective Search Space Reduction for Complex Question Answering over Knowledge Bases. *Proceedings Of The*

*Fifteenth ACM International Conference On Web Search And Data Mining*. pp. 172-180 (2022), <https://doi.org/10.1145/3488560.3498488>

- 7 Allen, R. Lost and Now Found: The Search for the Hidden and Forgotten. *M/C Journal*. **20** (2017), <https://journal.media-culture.org.au/index.php/mcjjournal/article/view/1290>


## 4.2 Current Practices of Iterative Classification Approaches for Digitised Historical Newspaper Collections

*Mickaël Coustatsy (University of La Rochelle, FR)*

*Estelle Bunout (University of Luxembourg, LU)*

*Jana Keck (German Historical Institute Washington, US)*

*David A. Smith (Northeastern University – Boston, US)*

License  Creative Commons BY 4.0 International license  
© Mickaël Coustatsy, Estelle Bunout, Jana Keck, and David A. Smith

The objective of this working group was to bring together scholars from different disciplinary backgrounds (Digital History, Computational Literary Studies, Natural Language Processing, and Computer Vision) to collect and compare current practices of iterative classification processes for historical newspaper collections. Current work on classification covers several tasks ranging from improving OCR or OLR to enriching semantic annotation of newspaper texts. These methods of classifier refinement, however, do not take advantage of the structure of historical newspaper collections, the punctuated equilibrium of newspaper layout, the evolution of language, the spatially distributed information cascades that spread news and other cultural artefacts. We propose, therefore, a process of structurally-informed exploration as important for building historically useful classification systems.

**Digitised newspapers enable a variety of classification tasks.** The digital turn in the study of historical newspapers and other sources rests on the creation of digital metadata and digital editions. At the lowest level, this includes digital photography, automatic image classification, and automatic transcription of text via optical character recognition. Recent work applying models from machine learning and artificial intelligence to historical sources, therefore, builds on the output of earlier digital work. Lara Putnam, in her essay on “The transnational and the text-searchable” [7], distinguishes this prior phase of computational research as the digitised turn, which receives less theoretical attention due to its congruence with (current) research habits. As Putnam writes, “*Precisely because web-enabled digital search simply accelerates the kinds of information-gathering that historians were already doing, its integration into our practice has felt smooth rather than revolutionary. . . How can typing words into a search box – which feels as revolutionary as oatmeal – be a sea change?*”. The digital format opens up these collections to the detection of patterns [2] that can be significant for humanist research, such as Long *et al.* [9] have shown with their study of haikus published in the US press. There have been contributions from the (digital) media history that show how text mining can help identify and discuss the emergence and distribution of rubrics in the press, using e.g. classifiers [3]. These few examples illustrate the exciting potentials the combination of digitisation and text mining tools offers; these research outputs, however, remain too often inaccessible for further uses by the community of (digital) humanists, hindering a deeper engagement with the source material beyond keyword search.

**Current classification approaches for historical newspapers are numerous.** At first glance, they seem very different depending on the disciplinary background: scholars use, for instance, search queries to explore documents and categorise them into relevant or non-relevant items to study events; they also use supervised machine learning to classify documents into newspaper genres to examine how genres have developed over time; or they combine textual and visual embeddings to group different items of a newspaper page (e.g. image and text) that belong together to enhance segmentation. As this shows, digitised newspapers are a very special kind of source given their abundance, heterogeneity, or seriality, and people from different backgrounds approach them from completely different perspectives. These differences have an impact on how scholars construct classification processes. Therefore, classification approaches for historical newspaper collections need to be reflected upon from a transdisciplinary perspective, considering the steps that precede the classification.

What all these multi-level classification approaches have in common is that they aim to improve the usability, searchability, and analytic capabilities for studying news of the past. Documenting these steps are necessary to share them with different disciplines. At the same time, reflecting upon these research projects brings to the surface the flaws of digitised collections and digital archives. However, this information is relevant and can be used as feedback for institutions that are digitising material and making it available online. Especially as in the case of digitised newspapers, two other idiosyncrasies make the findings more difficult to interpret for humanist researchers: the colossal size and the structure of each issue (itself being subject to historical changes). Searching for a word can generate interesting connections between contexts of its uses. For instance, looking for “morphine” in Dutch newspapers in the 19th century leads to hits in rubrics of hard news containing medical topics but also in fiction [8]. And this contextual information is very useful to interpret the results, not only to understand the distribution of the word, but also to reflect on the intertwining of themes across article types at a given time.

**Classification is a research output.** Current work on classification in historical newspapers covers several tasks. Layout analysis and page segmentation determine article breaks, figures, and reading order. Genre classification helps researchers to cluster articles. Entity linking connects mentions of named entities or definite descriptions (e.g., “the present king of France”) to entries in a knowledge base (where such entries exist; see Section 4.1). Although large pretrained language models and image representations have improved many tasks in natural language processing and computer vision, the main paradigm for applying machine learning to classification in historical newspapers is supervised machine learning. Researchers wishing to train a supervised classifier must annotate items – e.g., pages, or lines, or articles, or names, depending on the task – with labels indicating their class. But to perform this labelling, a researcher or research team needs to agree on what classes they are interested in. Determining a useful classification scheme for a research project, we propose, is closer to the information-seeking process researchers employ when formulating search queries and combing through results than it is to the item-wise labelling used in training a simple classifier. Importantly, researchers engaged in search do not usually stop with the results of one query; instead, they often reformulate a query to see if it returns more useful results or to answer some subsidiary question. In addition to these general issues with classification, the variation of historical materials across space and time makes it important to reason not only about our classifier’s average error rate but also about other properties of the error distribution. In this report, we explore how existing and speculative search capabilities for digitised historical newspapers can support the development of classifiers and classification schemes.

**Keyword search as a classification task.** Current research infrastructures for digitised newspapers supports some possible first steps in a classification workflow. Many digital collection search platforms allow users to assign labels to individual items or to save groups of items or search results in collections or work sets. Many platforms offer similarity-based recommendations, linking items (books or articles) to similar ones (“more like this”). Some of these similarity recommendations use only metadata, others use low-dimensional representations (e.g., embeddings or the output of black-box classification algorithms) to compute item similarity. Unfortunately, we are not aware of any bibliographic or newspaper search interfaces that allow user control over these representations for similarity search. To this functionality for supporting classification, we should also add keyword search. When users specify keywords and phrases, they are already engaged in a process of model building. Of course, until they examine the results, they may be, in Nicholas Belkin’s formulation [4], in an “anomalous state of knowledge”. If you know exactly how some concepts would be described in a historical document, then you would already know most of what you were searching for in that collection; because you need to search to fill gaps in your knowledge, you will inevitably not be able to describe the concept fully. This concept slippage or “vocabulary mismatch” is often easier to see with historical distance, as when, for example, nineteenth-century descriptions of reproductive health or infectious diseases do not match current terminology.

We can therefore think of a user query to a search engine as an incomplete and poorly calibrated model of the concept the user wants to find: incomplete, because it often contains only a few words or phrases representing the concept; and poorly calibrated, because humans are bad at estimating probabilities and assigning quantitative importance to the constituent terms of a query. Once users have obtained the first results of a query, they can however modify the original “model” by adding, deleting or modifying search terms. When performed automatically, this modification of the original query in response to users’ judgements of document relevance is termed “relevance feedback”.

**A call to share the classification models applied to digitised newspapers.** When users have labelled some initial examples – either by manual annotation or as the result of a search query – we can see several possibilities for model refinement that will be easily achievable in the short term. First, we mention supervised training, the focus of many current attempts to apply machine-learning methods to historical newspapers. There is scope for feature engineering, prompt engineering, and model selection, depending on the model architecture chosen. Improvements in creating annotated training and test sets have also been the focus of several projects. Secondly, and less widely used, has been query expansion through relevance feedback. Creating the model via manipulation of a human-readable query has some advantages. Finally, similarity search is supported by many newspaper platforms at the level of individual items, but not at the level of sets of items. Especially for those systems that compute similar items by projecting text or images into a low-dimensional embedding space, suggesting items that are slightly farther away than the nearest neighbours might improve the recall of concept expansion.

These methods of classifier refinement, however, do not take advantage of the structure of historical newspaper collections: the punctuated equilibrium of newspaper layout, the evolution of language, the spatially-distributed information cascades that spread news and other cultural artefacts. We propose, therefore, a process of structurally-informed exploration as important for building historically useful classification systems. For instance, when collecting training data for page layout models, we should sample a full range of historical periods for the newspapers of interest. Alternatively, we could have users check the predictions of a trained layout model over a broad temporal range.

**Classification as part of digital source criticism.** Interpretative work with digitised newspapers is mediated through image processing, text transcription, and search technologies. In addition to the filters of who and what gets recorded and archived, we observe differences in the effectiveness of optical character recognition, image analysis, and document classification. If these archival and digital filters removed or corrupted data uniformly at random, they might not affect our analyses, but they are often correlated with variables of interest, such as document date. To take a simple OCR example, the *Chronicling America* portal to the data from the US Digital Newspaper Program contains all issues of the *Richmond Daily Dispatch* from 1852 until its change of name in 1884. The word ‘Virginia’ appears at least once on 96% of these pages, but this average conceals an uneven trajectory over time. Starting in 1880, “Virginia” appears on only 84% of pages, compared to 98% before that time. The beginning of 1880 also corresponds to new microfilm rolls and a new batch (`vi_journey_ver01`) in the digitisation workflow. Comparing the *Daily Dispatch* before and after the beginning of 1880 falls prey to this time-dependent distortion. Although a recall of 84% in these noisier issues might still be useful, we are less able to generalise about the prevalence of terms. If, for example, we are interested in the rise of “scientific” racism in the 19th century and search the *Dispatch* for ‘Caucasian’ (as used in such anti-Reconstruction organs as the *Lexington [Missouri] Weekly Caucasian*), we find it on 0.68% of pages before 1880 and 0.33% thereafter. Should we conclude that “Caucasian” was used only half as often after 1880?

We observe similar effects with more complex problems. In the Newspaper Navigator experiments reported by [5], for example, page-element classifiers based on image features achieve an average precision of 74% at detecting headlines in *Chronicling America* pages from the first quarter of the twentieth century but 52% for 1875–1900 and 21% for 1850–1875. Similar differences in classification accuracy were observed for advertisements, illustrations, and other categories. These models are still useful for many applications – just as many other retrieval systems can still be useful at 20% average precision – but varying accuracy over time makes it more difficult to answer questions about changing page layout, the advertising basis for newspaper publication, and other topics. These errors analysing images and transcribing text in historical newspapers arise from mismatched training sets and data shifts.

At these error rates, it is difficult to ensure that any individual word or document is correctly classified, just as an inaccurate medical test may lead to improper decisions in particular patients’ cases – and just as machine learning can cause harm in other domains when applied to individuals. But in epidemiology, as in many social scientific and historical investigations, we can frame questions about the prevalence of a particular disease (or behaviour, or linguistic feature), even if we remain unsure of any given individual’s medical state. In a classic result from epidemiology, Levy *et al.* [6] showed how to derive unbiased estimates of the proportions of a population falling into two classes (e.g., infected and not infected) given noisy tests. We need to know the test sensitivity – i.e.,  $p(\hat{T} = 1 | T = 1)$ , the probability that a positive case will be correctly detected – and its specificity – i.e.,  $p(\hat{T} = 0 | T = 0)$ , the probability that a negative case will be correctly detected. If the proportion of the population whose tests are measured as positive is  $p(\hat{T} = 1)$ , then the corrected estimate for the proportion of positive cases is

$$p(T = 1) = \frac{p(\hat{T} = 1) - [1 - p(\hat{T} = 0 | T = 0)]}{p(\hat{T} = 1 | T = 1) - [1 - p(\hat{T} = 0 | T = 0)]}$$

**Combining classifications to mitigate their individual limitations and explore digitised newspaper collections.** If we train a classifier to estimate the proportions of a document collection falling into various classes, we can use information on the error distribution of

this classifier to correct these estimates. In an example of search in noisy OCR, assume for now that specificity for most queries is nearly 1 – i.e., it is very unlikely that one long word would be corrupted into a different long word. Further assume that sensitivity scales with the length of the query word. If we estimate the character error rate for the *Richmond Dispatch* as 10% before 1880 and 20% thereafter, the corrected percentage of pages with “Caucasian” before 1880 would go from 0.68% to 1.8%, and for pages from 1880 from 0.33% to 5.1%, suggesting this term’s frequency continued to increase.

Other possible directions for structurally-aware exploration include:

- Analysing the variation in classifiers and representations trained on multiple datasets as input;
- Speculatively exploring the consequences of alternate annotations on classifier predictions, toggling between document-level, feature-level, and corpus-level predictions; and
- Checking our understanding of what the classifier is learning by generating synthetic data, e.g., using a language model to generate new examples of a genre.

**Where to share classifications of digitised newspaper content?** Classification activities are dependent on existing infrastructures since they build on previous digital work for building input document representations, features, and annotations. We expect that many collections of digitised newspapers will provide application programming interfaces (APIs) not only for accessing images or OCR transcriptions of individual pages, articles, issues, or other metadata, but also for submitting search queries and retrieving results. To describe how these APIs might support classification workflows, we can distinguish at least five general kinds of queries they accept:

- unweighted keyword search, possibly with boolean and phrase operators;
- weighted keyword search, where individual terms, phrases or predicates may be assigned weights in the relevance function;
- dense text embedding retrieval, which takes a fixed-dimension vector representation of a query and returns documents or passages by their similarity to this vector;
- dense image embedding retrieval, which takes a vector representation of (part of) an image and returns (parts of) images; and
- document similarity retrieval, which accepts a single document as a query and returns other documents.

These query types do not exhaust the space of possible retrieval systems for digitised periodicals, as illustrated by the range of capabilities in *impresso*<sup>30</sup>, *NewsEye*<sup>31</sup>, and others. They do, however, form a useful set of primitive operations supported by several of these platforms.

Many document classification systems achieve acceptable accuracy using bag-of-words models with linear decision functions (e.g., logistic regression). We could thus retrieve likely members of particular classes using weighted keyword search and perform online updates to our model using relevance feedback. Unweighted keyword search would require more scaffolding on top of methods such as decision lists and random forests, but could be optimised end-to-end.

Several libraries have also experimented with similarity search by encoding text and image data using large pre-trained neural models. After mapping, say, an image into a fixed-dimensional vector space, these systems then perform a nearest-neighbour search in that

<sup>30</sup> <https://impresso-project.ch/app/>

<sup>31</sup> <https://www.newseye.eu/>

space to retrieve similar images. Where these vector-similarity searches could be exposed via an API, they can form useful primitives for classification systems. In some cases, a retrieval system will encode documents using a known published model such as BERT-BASE-UNCASED. We could directly encode queries using this same model. If the document-encoding model is not known, or if we wish to improve on this baseline performance, we could train our classification systems to learn improved query encoders.

In some cases, a retrieval API might only return nearest-neighbours for individual items in the collection rather than allowing arbitrary vectors as queries. In addition to supporting fast approximate relevance feedback [1], this clustering information can form the basis for a classifier.

## References

- 1 Cartright, M.-A., Allan, J., Lavrenko, V., McGregor, A. Fast query expansion using approximations of relevance models. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)* (2010)
- 2 Dzogang, F., Lansdall-Welfare, T., Team, F. N., & Cristianini, N. (2016). Discovering Periodic Patterns in Historical News. *PLOS ONE*, 11(11), e0165736. <https://doi.org/10.1371/journal.pone.0165736>
- 3 Langlais, P.-C. (2022). Classified News. Revisiting the history of newspaper genre with supervised models. In E. Bunout, M. Ehrmann, & F. Clavert (Eds.), *Digitised Newspapers – A New Eldorado for Historians?: Tools, Methodology, Epistemology, and the Changing Practices of Writing History in the Context of Historical Newspaper Mass Digitisation*. De Gruyter Oldenbourg. <https://www.degruyter.com/document/isbn/9783110729214/html?lang=en>
- 4 Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, 5(1), 133–143.
- 5 Lee, B.C.G., Mears, J., Jakeway, E., Ferriter, M., Adams, C., Yarasavage, N., Thomas, D., Zwaard, K., Weld, D.S. The Newspaper Navigator Dataset: Extracting And Analyzing Visual Content from 16 Million Historic Newspaper Pages in Chronicling America. <http://arxiv.org/abs/2005.01583> (2020)
- 6 Levy, P.S., Kass, E.H. A three-population model for sequential screening for bacteriuria. *American Journal of Epidemiology*, 91(2):148–154 (1970)
- 7 Putnam, L. The Transnational and the Text-Searchable. *American Historical Review*, 121(2):377–402, (2016)
- 8 Walma, L. W. B. (2015). Filtering the “News”: Uncovering Morphine’s Multiple Meanings on Delpher’s Dutch Newspapers and the Need to Distinguish More Article Types. *TS: Tijdschrift Voor Tijdschriftstudies*. <http://dSPACE.library.uu.nl/handle/1874/324205>
- 9 Long, H., & So, R. J. (2015). Literary Pattern Recognition: Modernism between Close Reading and Machine Learning. *Critical Inquiry*, 42(2), 235–267. <https://doi.org/10.1086/684353>



### 4.3 Fairness and Transparency throughout a Digital Humanities Workflow: Challenges and Recommendations

*Kaspar Beelen (The Alan Turing Institute – London, GB)*

*Sally Chambers (Ghent University, BE & KBR, Royal Library of Belgium, Brussels, BE)*

*Marten Düring (Luxembourg Centre for Contemporary and Digital History, LU)*


*Laura Hollink (CWI – Amsterdam, NL)*

*Stefan Jänicke (University of Southern Denmark – Odense, DK)*

*Axel Jean-Caurant (University of La Rochelle, FR)*

*Julia Noordegraaf (University of Amsterdam, NL)*

*Eva Pfanzelter (Universität Innsbruck, AT)*

**License**  Creative Commons BY 4.0 International license

© Kaspar Beelen, Sally Chambers, Marten Düring, Laura Hollink, Stefan Jänicke, Axel Jean-Caurant, Julia Noordegraaf, and Eva Pfanzelter

#### 4.3.1 Main challenges and aim

How can we achieve sufficient levels of transparency and fairness for (humanities) research based on historical newspapers? Which concrete measures should be taken by data providers such as libraries, research projects and individual researchers? We approach these questions from the vantage point that digitised newspapers are complex sources with a high degree of heterogeneity caused by a long chain of processing steps, ranging, e.g., from digitisation policies, copyright restrictions to the evolving performance of tools for their enrichment such as OCR or article segmentation. Overall, we emphasise the need for careful documentation of data processing, research practices and the acknowledgement of support from institutions and collaborators.

Increasingly, historical newspaper data undergoes automatic processing using probabilistic methods. For example, topic modelling may inspire the identification of semantic facets within a set of articles, and word embeddings can suggest new keywords and as such different contexts or semantic shifts over time. The acknowledgement of such input matters inasmuch as it holds novel analytical potential and constitutes opportunities to broaden researchers' views on their sources. At the same time, it can mislead researchers due to the underlying principles which govern their creation and make them neither neutral nor objective. We therefore emphasise that researchers benefit from accessible information regarding the processing of data and its fairness. Still, at some point they will nevertheless have to trust systems' output and accept that their findings also depend on factors beyond their understanding, e.g., the impact of different constellations of search engine settings or the outcome provided by topic modelling tools.

Our goal is to compile recommendations for different aspects of transparency and fairness required for the analysis of digitised and enriched historical newspaper collections. We focus on aspects with a potentially high impact on the outcome of research. We distinguish between the need of researchers to obtain information for processes which lie beyond their control, such as institutional digitisation policies and OCR, and their obligation to provide information on aspects they can control, such as the documentation of their *modus operandi* and sharing research data to allow the traceability of their research. In this report we focus on the former.

The authors of this report have backgrounds in computer science (AI, visualisation, engineering), history (media history, contemporary history, digital history) and library science. This report is the result of one week of exchange and discussion on the topic of data transparency and fairness.

### 4.3.2 Approach

In a first exploration phase we started with a round-table discussion about fairness and transparency in the context of humanities research based on digitised historical newspapers. For a more formal and systematic review of interface features for historical newspapers see [4, 15].

Second, we performed an initial exploration of seven portals that provide access to historical newspapers. Several issues related to fairness and transparency surfaced in the round-table discussion and in the platform exploration which was centred on the needs of researchers in the historical disciplines.

Third, we used the output of the exploration phase to identify six focus areas which play a key role for historical newspaper research and formulated accompanying recommendations for measures to improve transparency and fairness. The focus areas and measures are organised along the lines of a typical digital humanities workflow.

In a final application phase we used the identified focus areas and recommendations to evaluate the *impresso* interface<sup>32</sup> which was developed with particular attention to transparency. We tested the portal and discussed to what extent each issue plays a role, and to what extent *impresso* implements or enables the recommended strategies. This resulted in insights regarding how far one of the state-of-the-art portals is when it comes to facilitating fair and transparent research on digitised historic newspapers.

### 4.3.3 Definitions

**User** Various types of persons work with digitised historical newspaper data, for example humanities scholars, interested lay people, collection owners, and portal developers, as well as scientists from other fields, such as natural language processing (NLP) researchers, who use newspapers as training sets. In this report, our point of reference are foremost the needs of historians, but we nevertheless expect that our recommendations are also relevant for other user groups.

**Collection** A comprehensive body of materials, in our case digitised historical newspapers, that is curated by a library, museum, or archive.

**Corpus or Research Dataset** The dataset that a researcher has compiled and on which they will do their analysis. The research dataset may be a subset of one or more collections. The researcher may have used one or more portals to compile the research dataset. The research dataset has often undergone multiple (iterative) processing and enrichment steps.

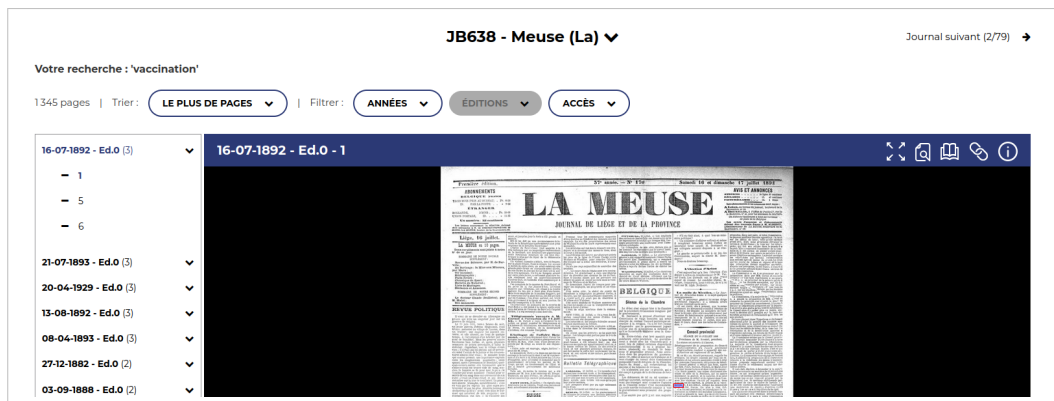
**Portal** An access point to one or more collections of digitised historic newspapers, providing functionality such as keyword search, faceted search, browsing or the inspection of raw scans.<sup>33</sup>

**Workflow** A sequence of actions executed by a researcher to find, collect, transform, enrich, and/or analyse documents.

**Fairness** We define fairness as the absence of bias. Fairness and equity can relate to a collection, a corpus, or the input and output of a tool. Fairness can be improved by raising awareness of biases as well as unwanted over- and under-representations. Lacking fairness can either be the result of culturally ingrained biases or technical processing on any stage of newspaper digitisation and enrichment.

<sup>32</sup><https://impresso-project.ch/app/>

<sup>33</sup>In this report the terms *platform*, *interface*, *web application* describe the same thing.



■ **Figure 8** Screenshot of the KBR Belgica Press search interface (© KBR, Royal Library of Belgium).

**Transparency** Explicit, accessible information regarding the content of a collection or corpus regarding the workflow that was followed to create, process, and enrich it and/or regarding what is known about its fairness.

#### 4.3.4 Exploration: Initial use case-based exploration of platforms

Here we present the findings of our initial exploration of platforms that provide access to historical newspapers. The findings were used as input for the workflow requirements regarding fairness and transparency described in the next section. Seven platforms were investigated. This list is not complete: not covered are, for example, Delpher<sup>34</sup>, the CLARIN Newspapers Resource Family<sup>35</sup>, and *impresso*.

The initial exploration was guided by a use case on the topic of “vaccination”. We have documented this exploration in the form of short reviews which are structured as follows:

- Overview of the portal and its collections including a characterisation of the titles and main features for search, exploration and opportunities to interact with the data.
- Vaccination case study with a focus on the following questions: When was the first article which mentions vaccination published? Which bursts/peaks can be observed in the coverage?
- Summary and assessment of the level of transparency and fairness.

In the following sections we provide reports on the results of these experiments for different portals.

#### BelgicaPress

The landing page of the Belgica Press portal<sup>36</sup> (Figure 2) gives information about the content of the available collection: 121 titles published between 1814 and 1970. Some details are given about the selection of this collection, as well as the information that only one title has been digitised until 1970. However, there is no further information concerning the availability of other titles.

<sup>34</sup> <https://www.delpher.nl/>

<sup>35</sup> <https://www.clarin.eu/news/clarin-resource-families-newspaper-corpora>

<sup>36</sup> <https://www.belgicapress.be>

The interface itself is simple. A search bar can be used to query for keywords and an advanced search allows for date filtering as well as Boolean conditions on the presence or absence of keywords in the results. A first query for “vaccin” yields 12.721 pages in 93 newspapers. The results are grouped by newspaper title which makes the search for the first occurrence and the overall distribution over time within the entire corpus rather laborious. Copyright-protected content is accessible for registered researchers with a MyKBR account. The results are presented as a list of newspaper titles sorted by the number of pages containing mentions of the keywords. When clicking on a result, a new page opens with a viewer allowing the user to see mentions of the keywords. The user can navigate through a list of other pages of this title. It is also possible from this page to switch to another title. There is apparently no relevance ranking for search results but there is the possibility to sort results by newspaper title, by date or by number of pages containing a keyword.

### German Newspaper Portal

The German Newspaper Portal<sup>37</sup> has a very simple, “clean” interface that is available in German and English. It is not immediately clear if the search is also bilingual; testing reveals that this is not the case. The caption on the search page has a very minimal indication of the scope of the collection: one can search newspapers from 1671 to 1950. The first thing users see is a search box which invites for a direct keyword search. If users scroll down, three different browsing options are provided. Underneath those is a graph visualising the total amount of newspapers. At the bottom there is a display of a historical newspaper issue of the same date 100 years ago. The interface is clearly designed for a general audience, that is: users focused on encyclopedic use and browsing.

The “About” page indicates that it is a federated site that provides access to newspapers held at different German institutions. It provides data on the total number of newspapers: “The Deutsches Zeitungsportal was launched in October 2021 with 247 newspapers, 591,837 newspaper issues and a total of 4,464,846 newspaper pages from nine libraries. The offerings are being continually expanded and, in the long run, should comprise all digitised historical newspapers which are stored in German cultural and scientific institutions.” They also indicate that it is not a representative selection<sup>38</sup>, but how “not representative” it is, is not indicated. There is an alphabetical list of all the newspapers with information on their publication history, frequency of publication, and area of distribution.<sup>39</sup> Only 82% of the articles are full-text indexed, but it is unclear which parts of the collection it concerns. This makes it very hard to do source criticism on this collection.

A keyword query for “vaccination” in the search box generates a graph and result-list with snippets organised by titles: apparently 279 results from 26 June, 1802, until 5 June, 1950, were found. Results can be sorted by relevance, but it is unclear how that is defined. Alternatives are sorting functions by publication date (oldest first, newest first) or A-Z or Z-A, where results are apparently ranked by newspaper title.

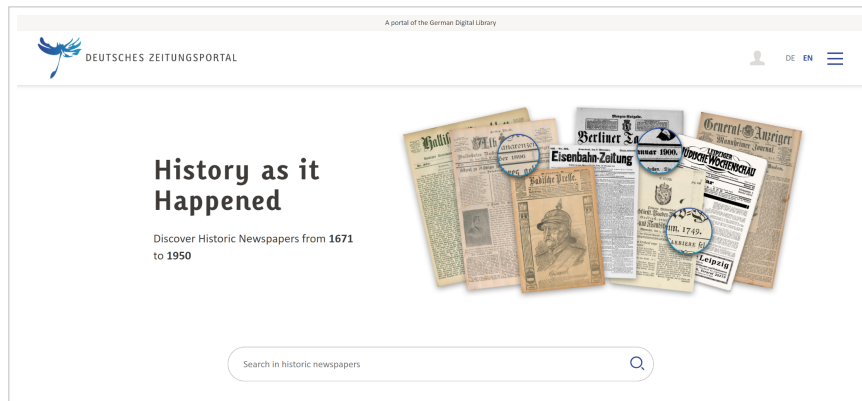
The earliest mention of “vaccination” is in the *Hallesches Tageblatt* of 26 June, 1802, where it is mentioned in a section on “Kuhpocken” (“cow pocks”).

However, a wildcard search of “vaccin\*” gives 759 results with the oldest in the *Gülich und bergische wöchentliche Nachrichten* of 20 May 1783, but there it mentions the Latin

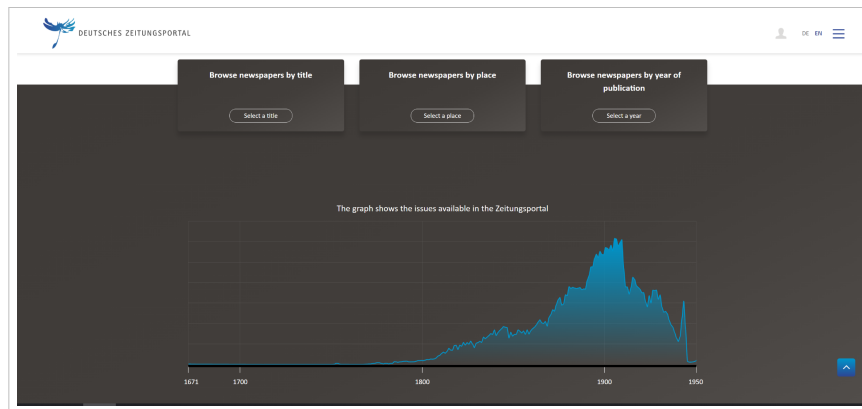
<sup>37</sup> <https://www.deutsche-digitale-bibliothek.de/newspaper>

<sup>38</sup> <https://www.deutsche-digitale-bibliothek.de/content/newspaper/fragen-antworten>

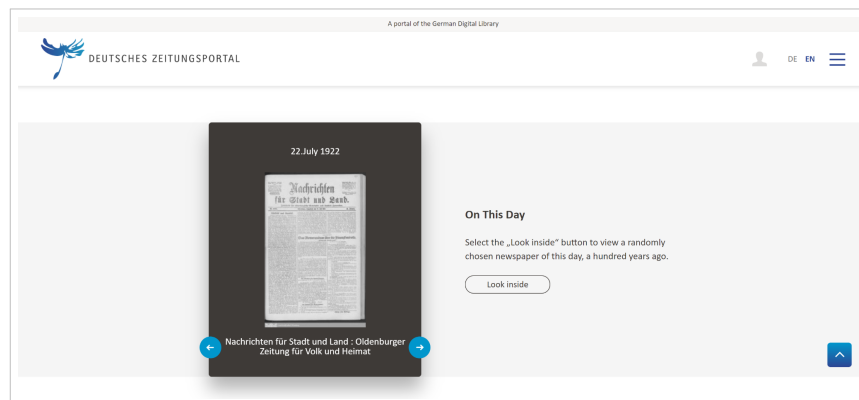
<sup>39</sup> <https://www.deutsche-digitale-bibliothek.de/newspaper/select/title>



(a) Search landing page of the Deutsches Zeitungportal (© DDB, Deutsche Digitale Bibliothek).

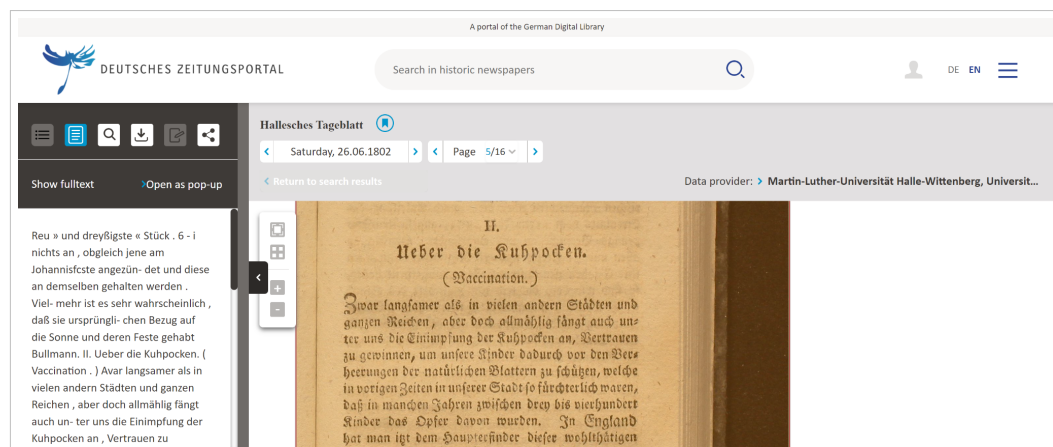


(b) Timeline showing the newspaper issues by year on the Deutsches Zeitungportal (© DDB, Deutsche Digitale Bibliothek).



(c) Example newspaper page on the Deutsches Zeitungportal (© DDB, Deutsche Digitale Bibliothek).

■ **Figure 9** User interface of the Deutsches Zeitungportal.



■ **Figure 10** Earliest example of “Vaccination” as shown on the Deutsches Zeitungsportal (© DDB, Deutsche Digitale Bibliothek).<sup>40</sup>

“Vaccinium” which refers to a blueberry<sup>41</sup> – considering that the word vaccination was invented by Jenner in 1796 this result clearly is off topic. The earliest mention from 22 February 1802, is in the *Karlsruhe Zeitung*, the newspaper that most often contains the term (107 articles, 14% of the total). The results page contains a result hit timeline that reveals peaks in 1871-1874 (coinciding with the smallpox pandemic of 1870-1874), 1884 (perhaps a late response to Pasteur’s publication on vaccination of 1880?), one around 1890 (perhaps new vaccinations found) and a final one in 1913 (with reference to the use of vaccinations at war time), after which the references decline.

The portal allows users to filter by newspaper title or distribution area (or period), but the functionalities are too limited for putting together a research corpus for our question. The FAQ section points to the well-documented API<sup>42</sup> where the portal allows digitally literate and registered users to extract data.

To conclude: the portal allows for exploratory search but is not suited for building a research corpus due to a lack of transparency on the scope and quality of the underlying collections and their processing. The API should be used to extract a corpus and for quality assessments, but this requires technical expertise most historians do not have.

### Europeana Newspapers

The Europeana portal includes a “Newspapers Theme”<sup>43</sup>. The title of the theme is “Explore the headlines, articles, advertisements, and opinion pieces from European newspapers from 20 countries, dating from 1618 to the 1980s.” It includes 887,607 items from ten European countries (Austria, Estonia, Finland, Germany, Italy, Latvia, Luxembourg, The Netherlands, Poland and Serbia). However, it is not possible to see a listing of which titles are included.

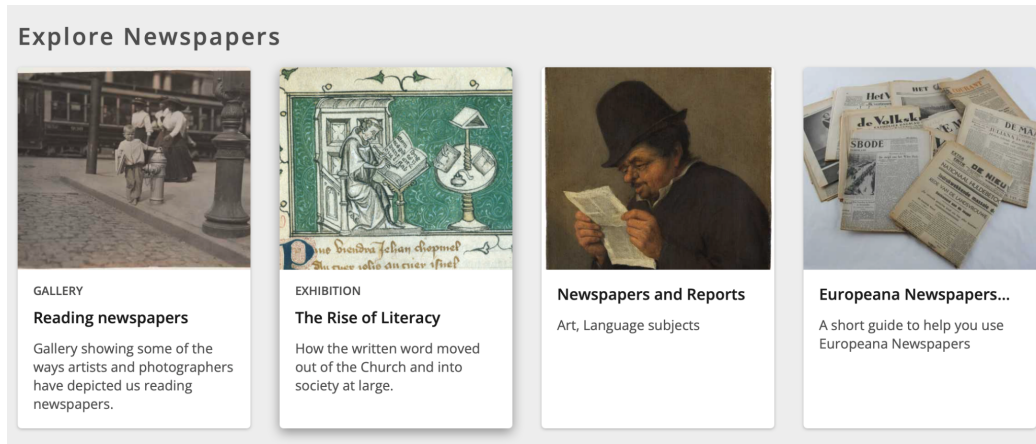
<sup>40</sup> <https://www.deutsche-digitale-bibliothek.de/newspaper/item/N5UFN05HCR36P7BJK3TYEI5XM4IR6UUK?issuepage=5>

<sup>41</sup> <https://www.deutsche-digitale-bibliothek.de/newspaper/item/R2LPDFW7YX27WTEOLNLEBIE4PCDKF66Q?issuepage=4>

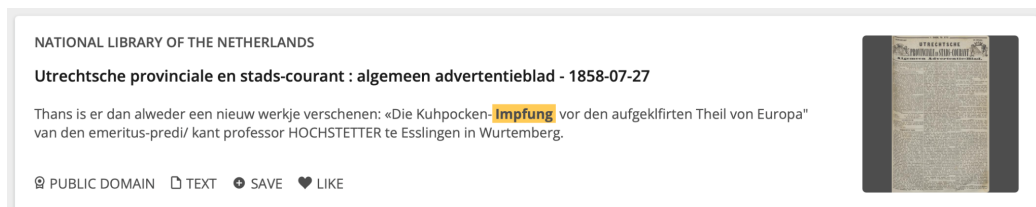
<sup>42</sup> <https://labs.deutsche-digitale-bibliothek.de/app/ddbapi/>

<sup>43</sup> <https://www.europeana.eu/en/collections/topic/18-newspapers>

Additional content is provided at the end of the page, including a gallery on Reading Newspapers, Exhibition on the Rise of Literacy, teaching information on Newspapers, Reports, and a short guide to the use of Europeana Newspapers.



(a) Landing page of the Europeana newspapers portal (© Europeana).



(b) Example of search results for the query “vaccination” as shows on the Europeana newspaper portal (© Europeana).

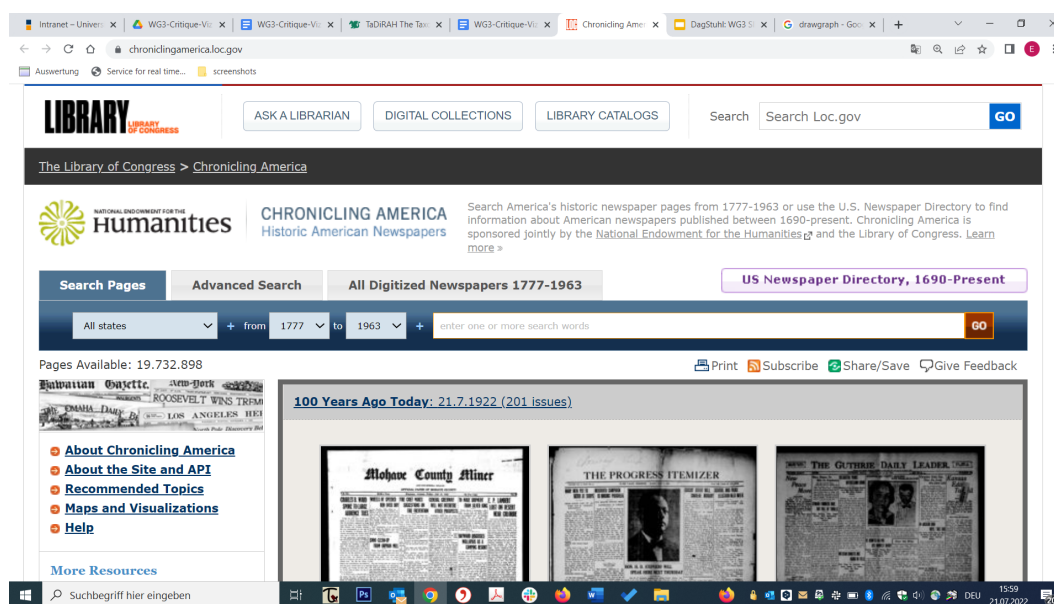
■ **Figure 11** User interface of the Europeana newspaper archive.

For the “vaccination” research questions, since it is a multilingual portal it is important first to assess which search terms could be used to find relevant newspaper articles on vaccines. Based on this author’s language skills a search was undertaken on “vaccine” (EN: 4,626 results), “Impfung” (DE: 12,647 results) and “vaccin” (FR: 4,607 results). Using a wildcard, e.g. “vaccin\*” (10,734 results), articles in other languages (e.g., Italian) were included. It was not possible to sort the results. However, a range of result filters are available, e.g., by language or providing country. Additionally there was a “date issued” filter which, however, was not easy to use. Each of the search results provided an overview of the providing institution, the title of the newspaper, as well as a text snippet including the search term, and a thumbnail of the newspaper in question (see below).

Relevant articles can be saved by the user in a personal “gallery” (following the creation of a Europeana account). This gallery can be kept “private” or made “public”. It is possible to share a public gallery on social media, e.g. on Twitter<sup>44</sup>. It is possible to download individual search results as page images (jpeg). There does not seem to be an advanced search function. There is some filtering of search results, however, they are not sufficient enough to answer our research questions.

<sup>44</sup> <https://www.europeana.eu/en/set/7143>





■ **Figure 12** Screenshot of the landing page of Chronicing America (© Library of Congress).

### Chronicing America

The landing page of the Library of Congress collection of historical US-newspapers<sup>45</sup> offers several facets and has tabs to give access to the collection and offers links to information pages, APIs, as well as help files, thematic corpora, maps, and visualisations. A search bar and tabs in the background indicate that there are more search options available for advanced search and more complex investigations of the collection. The attention of users is drawn to the centre of the page where a selection of newspaper front pages are displayed under the heading “100 Years Ago Today”, today’s date and the number of newspaper issues collected.

The collection is a composition of “historic US-newspapers from 1690 to the present”. The interface is the result of a collaboration between the Library of Congress and the National Endowment for the Humanities. It includes 3,758 newspapers with 19,7 million digitised pages. The APIs<sup>46</sup> enable expert users to perform the following tasks: search, auto-suggest from newspaper titles, link to stable URLs, linked data views of the collection, JSON view of data, bulk data to use with external services, and CORS- and JSONP-support for JavaScript applications. For all APIs explanations on use and examples are given. Under the heading “Recommended Topics” thematic features in Chronicing America are collected. These corpora are arranged alphabetically, by category, and by date range. They cover a growing number of different themes, time-spans, and genres. The section heading “Maps and Visualizations” leads to a number of graphs and data visualisations of the collection. These pages are updated on a regular basis. So, while the landing page and the simple search bar may give the impression that this collection is meant for a general audience, both the sub-sites and the accessible design of the “Advanced Search” function are clearly intended for expert users.

<sup>45</sup> <https://chroniclingamerica.loc.gov/>

<sup>46</sup> <https://chroniclingamerica.loc.gov/about/api/>

The collection is composed of newspapers in 19 languages: English is the dominant language (with 18,7 mio pages), followed by German (500,000), and Spanish (330,000). At the end of the scale Hebrew (830) and Arabic (2,000) can be found.

With regard to transparency and fairness we wish to highlight dedicated visualisations on the distribution of ethnic press coverage within the corpus. A keyword query for “vaccine” using the basic search bar leads to 208,360 results. The wildcard search for “vaccin\*” to capture also results for “vaccination” or “vaccinated” led to slightly over 207,000 results. This apparently wrong output was quickly resolved by reading the help files which indicated that wildcards, as well as upper-/lower-case search, and simple Boolean operators are not implemented. However, the search engine utilises language specific dictionaries which use stemming to include word variants. In order to limit (or increase) the search results, combinations of words or the features offered in the “Advanced Search” should be used (here filters on states, titles, years, front pages, language, combination of words, phrase search, and distance search are implemented). The search for “vaccine fear” produces 69,762 results. A quick scan of the results showed, however, that the two terms often do not occur in the same news item so that this keyword search does not produce usable results. A distance search of the two terms (with a distance of 10 words) produced 1,344 results which did not prove more appropriate (corresponding to sentences similar to “I fear that . . .”). Finally, the combination of the terms “vaccine” and “effect” in a distance search of 10 words led to 5,634 results that could be used to study newspaper reporting of this topic. However, it remains uncertain if the word “effect” really covers what a user was looking for in the context of discourses on vaccination. Bulk downloads are not possible at this level. Another point of “granular access” (as opposed to bulk download) is the Chronicling America API. Programmatic access is often preferable for computational analysis, as retrieving and processing data can be easily integrated into one workflow. However, the API functionality is in many ways similar to keyword search. The main functionality is search defined by a query term and refined by a few additional parameters. As can be gathered from the online documentation, the API is especially useful when a researcher wants to retrieve documents related to a specific topic in bulk. Of course, additional filtering can happen downstream in custom-made scripts, but it does not seem to be part of the API functionality (or at least is not very well publicised on the main page).<sup>47</sup> Having said that, the API is undoubtedly easy to use and the examples are easily adaptable. We successfully used the search endpoint to retrieve articles that mention “vaccination” as a starting point for further processing.

## ANNO

AustriaN Newspapers Online (ANNO)<sup>48</sup> is a digitisation project of the Austrian National Library for Austrian historical newspapers and magazines. The project was launched with 15 newspapers in August 2003, and now, more than 25 million pages of more than 1,500 newspapers and magazines can be read and downloaded free of charge and in full text from the portal. The oldest editions date back to 1568. Like other newspaper portals, ANNO offers users to browse and read digital newspapers, and to search for articles based on keyword or an advanced search with additional filters (publication place, date, language, and topic). First hits for “vaccin” are found in the Italian paper *Il Corriere ordinario* from 1679, however referring to cows, a common false positive result we have also observed in other portals.

---

<sup>47</sup> <https://chroniclingamerica.loc.gov/about/api/>

<sup>48</sup> <https://anno.onb.ac.at/>

Medium	<
Titel	<
Erscheinungsort	<
Sprache	<
Zeitraum	∨
1731-1787	33
1788-1845	5.654
1846-1902	15.736
1903-1960	14.279
1961-2019	839
Thema	<

(a) Search results by period on the ANNO interface. The strategy on how the temporal facets shown on the left were defined is intransparent (© österreichische Nationalbibliothek).

Salzburger Zeitung 16. März 1744 ✕

1 von 1 Ergebnissen für "impfung" in dieser Ausgabe:

[Seite 2](#)

...phe durch die Fortpflanzung in ihrer ursprünglichen Kraft ein-  
 büße, wo hingegen **Impfung**, unmittelbar von der Kuh entnom-  
 men, die volle heilsame Wirkung äußere. Da die Ansteckung  
 ...phe durch die Fortpflanzung in ihrer ursprünglichen Kraft ein-  
 büße, wohingegen **Impfung**, unmittelbar von der Kuh entnom-  
 men, die volle heilsame Wirkung äußere. Da die Ansteckung...

(b) Screenshot of an individual result (© österreichische Nationalbibliothek).

■ **Figure 13** User interface of the ANNO portal.

ANNO does not provide visual cues that summarise metadata of the retrieved results such as, e.g., *impresso* does. The results are displayed in a faceted browser environment, and facets, for which numerical information are provided, can be selected and deselected. The default ranking of results is by relevance, however, it is not traceable how relevance is defined. Ordering of results by other metadata like date is also possible. Clicking a result opens a popup that juxtaposes scan and OCR transcript, and highlights the search term(s) in both views. ANNO also includes the option to use Boolean operators. Alongside this common search and filter features, ANNO supports filtering by language and themes such as “science” or “agriculture” but it remains unclear, how these filters and the underlying data were generated. A Help and FAQ sections explains available functionalities but do not include information about the technical processing.

### NewsEye

The NewsEye portal<sup>49</sup> includes newspaper data from various countries. The data are from different time periods, which makes a comparative analysis difficult. Keyword search in combination with filters can be used to search through the data. Results are presented as snippets, allowing to quickly assess the relevance of the results, and thus the appropriateness of the keywords. The portal allows users to create and store a custom research dataset by selecting articles or newspaper issues from the search results page. This increases transparency for peers with regard to which data a researcher used for their analysis. The portal does not support transparency with respect to the methods used to select data. A systematic method could be, for example, to fix a set of keywords/facets, and include the resulting articles.

The portal contains an interface to create and store experiments, i.e., sequences of data processing steps (Figure 14a). This functionality increases transparency of the analysis step: not only can the pipeline be stored for future use, it also makes it easy to compare output of different pipelines.

The NewsEye platform used for the current exploration was the experimental platform of the NewsEye project. So, some functionalities were not implemented in this interface yet: the help-button did not work yet; some of the facets still produced unexpected results; only a small number of simple data processing tools were included in the experiment interface (e.g., stopword removal).

The search for the truncated word “vaccin\*” produced 25,255 search results in Swedish, French, English, German, and Finnish (Figure 14b). A random check of documents in the different languages confirmed that these were indeed related to vaccination. This result is not surprising as in many languages the word vaccination was derived from the Latin “vaccinus” (from the cow). A graph gives an overview of the distribution of the search hits over time. Several facets allow researchers to dig deeper into the search results. Results can additionally be sorted by date or relevance score and the function “random sample” gives a quick overview of what the reader can expect to find in the results. The first mention of “vaccin\*” in the newspapers aggregated in NewsEye is in Swedish from the Finnish title *Abo Underrattelser* from 13 March, 1824, where the distribution of vaccines in Finland is discussed.

### Trove

The Trove portal<sup>50</sup> aggregates a wide range of textual and visual digitised and born-digital resources (books, newspapers, websites, images), hosted by Australian cultural heritage institutions. Trove Newspapers offers a clean interface with common (advanced) search and filtering options alongside the notebook-based GLAM-workbench<sup>51</sup>. An informative About<sup>52</sup> section gives a concise overview of the whole “Trove ecosystem”, its construction and guiding principles and is accompanied by a Research Guide<sup>53</sup> which offers basic insights into the availability and legal status of the collection. User expectations are managed effectively through additional documentation, e.g. on a variety of errors and instructions for correction

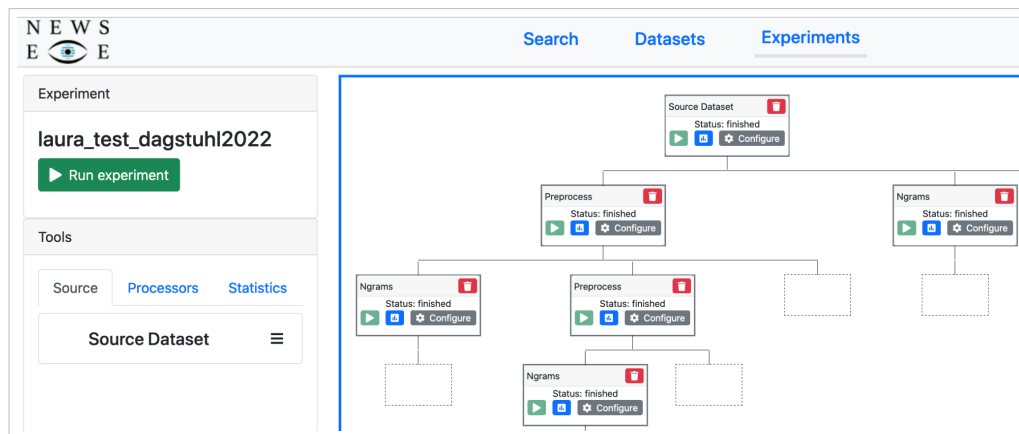
<sup>49</sup> <https://platform2.newseye.eu/>

<sup>50</sup> <https://trove.nla.gov.au/newspaper/>

<sup>51</sup> <https://mybinder.org/v2/gh/GLAM-Workbench/trove-newspapers/master?urlpath=lab/tree/index.ipynb>

<sup>52</sup> <https://trove.nla.gov.au/about>

<sup>53</sup> <https://www.nla.gov.au/research-guides/australian-newspapers>



(a) Experiment workflow on the NewsEye interface. Screenshot of NewsEye interface to interactively create and store experiments (© NewsEye).

The screenshot shows the NewsEye search results for the query 'vaccin\*'. The interface includes a search bar, a results list, and a sidebar with filters. The search results are as follows:

Result ID	Publication date	Newspaper	Snippet
1. hufvudstadsbladet 528326 article 166	1899-12-20	Hufvudstadsbladet	...Animal vaccin...
2. hufvudstadsbladet 528326 article 658	1899-12-20	Hufvudstadsbladet	...Stadsbarnmorska. Fru M. Rautell, Elisabetsgatan 12. Vaccination. 8 Övist Anareg. 15.10—11...
3. hufvudstadsbladet 1182590 article 680	1914-02-26	Hufvudstadsbladet	... lighet i refererat, men det kan vaccinerade och möjligen påförda utslutande ensidigt. Var det icke också...
4. abo_underrattelser 365838 article 9	1865-05-11	Abo Underrattelser	... skarpkbat. i Hfors. Resande. Maj 7. Handlne Söderström och Råstedt fr. Somero, samt vaccinator Finnberg fr...

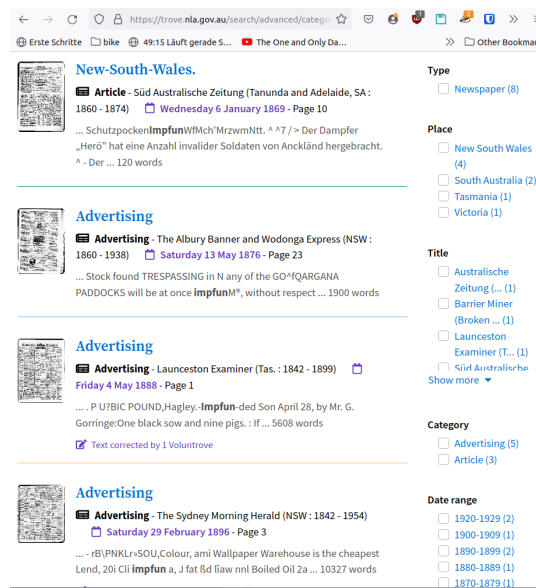
(b) Search results for “vaccin\*” in the NewsEye platform (© NewsEye).

■ **Figure 14** User interface of the NewsEye portal.

for volunteers.<sup>54</sup> Noteworthy is also optional information concerning cultural sensitivity which users are free to en- or disable throughout their interaction with the portal. During this limited testing we were however not able to see it in action but learned that users are encouraged to amend DublinCore and MARC metadata of affected articles with the reference “Culturally sensitive”. Users are furthermore able to filter content by region, content type, media, and content length with the notable absence of language.

Our case study on vaccination reveals the tremendous added value of crowd sourcing and its effective implementation in Trove. The earliest reference can be found with a query for “vaccin\*” and retrieves an article published in the *Sydney Gazette and New South Wales Advertiser* in 1803. The article covers an experimental treatment of orphans with early vaccines against cow pox including the assertion that “It is believed, that it never has been

<sup>54</sup> <https://trove.nla.gov.au/help/become-voluntrove/text-correction>



■ **Figure 15** Ranking of search results for “Vaccination” as shown in the Trove interface (© National Library of Australia and Partner Institutions).

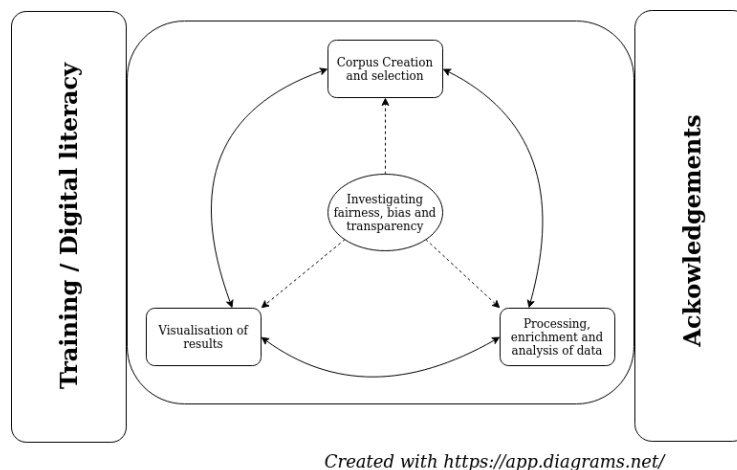
fatal, and never will be”. The query term “vaccin\*” does not occur in the text which has been manually transcribed by volunteers. Instead, the article has been tagged manually with the term “Vaccination” alongside other helpful yet anachronistic tags such as “Bioethics” or “Clinical trial”. Trove search includes such tags as well and thereby helped retrieve this article. A distribution over time of search results is possible via hits per year counts but this feature is basic. Information e.g. regarding the breadth of content type detection across the corpus is missing as is information concerning the overall representativity of the corpus for the Australian press. The interface supports crowdsourced OCR correction and informs about the number of “Voluntroves” who worked on a given article. Overall, Trove stands out regarding audience-integration: Crowdsourcing and -annotation features, notebook-infrastructure and cultural sensitivity are well integrated and cater to the needs of different user groups.

#### 4.3.5 Consolidation: Issues and recommendations with respect to fairness and transparency in each stage of the workflow

We identified distinct stages in a typical digital humanities workflow for the analysis of digitised historical newspapers:

1. Research corpus creation and selection,
2. Processing and enrichment,
3. Data analysis,
4. Visualisation of results,
5. Training,
6. Acknowledgements.

Figure 16 illustrates how the stages interrelate. The stages are often performed in an iterative fashion. In this section, we discuss issues with respect to fairness and transparency in each of the stages, and present recommendations for how to deal with them.



■ **Figure 16** Typical workflow of a researcher. It is essential to investigate fairness, bias and transparency at various stages and in a continuous fashion. It is to be noted that if the training is a prerequisite to the creation of a research dataset, it never really stops (© Axel Jean-Caurant).

#### 4.3.6 Research corpus creation, selection and sharing

##### ► Focus areas for transparency and fairness

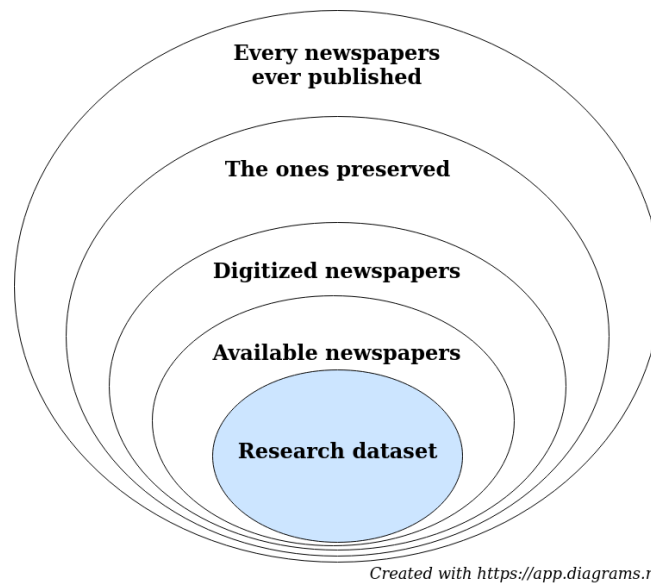
**Awareness of the digitisation and preservation policies.** The “digital sample” constitutes a subset of the totality or population of newspapers which existed at some point in time. This population is unknown, but can be approximated using contextual resources such as newspaper press directories or library catalogues [2]. Both are simultaneously useful and problematic. Catalogues record mainly what sources are preserved and are close to providing an overview of the complete newspaper record especially for countries with legal deposit. Newspaper press directories are a useful historical source, but also come with their own issues: classification of what “is” a newspaper changes over time and varies by directory. Ultimately, the population or “newspaper landscape” remains unknown, but we can describe those that have been recorded and/or preserved using metadata derived from catalogue or contextual resources [2].

A rich description of the newspaper landscape helps to contextualise and situate the “sample” of digitised newspapers. In other words, it enables us to at least approximate the “representativeness” of the collection, bearing in mind that the latter concept is more complex than simple proportionality and is always defined in relation to the research questions and ethical values or priorities of the researcher. When it comes to the composition of a corpus, we as researchers and content providers will never be able to get rid of biases and unwanted over- and under-representations. Our goal must rather be to identify, understand, acknowledge them, to infer how they may influence the research outcomes and to make them clearly visible within portals.

When using newspapers at scale and-to repeat the metaphor-as a “mirror” of the past, assessing diversity of collections emerges as a critical issue alongside the processes of media production which heavily influences how the presence was reflected. Researchers need to acknowledge whose perspectives or voices are absent in the data and which social categories dominate a collection.

The question of representativeness is closely intertwined with the issue of diversity and inclusion: which (social) perspectives are present in our data, which are missing? Coming back to the metaphor of newspapers as a “mirror” of the past, we can not simply trust the





■ **Figure 17** A research dataset is inherently biased, as it is impossible to create a complete dataset because of missing or unavailable sources (© Axel Jean-Caurant)

reflection but need to assess how it potentially distorts our image of the past. With rich descriptions of the sample (and population) we can situate data historically and socially. Of course these will always be rough and approximate descriptions, but nonetheless a crucial part of contextualising (the results derived from) big data. Finally, digitisation does not automatically mean accessibility which depends on the institutional policies (e.g. paywalls) and legal restrictions.

**Diverse user needs.** Keyword search may satisfy a large group of researchers (and laymen), but others may want to go beyond simply retrieving and reading newspaper content. Interfaces simultaneously provide and restrict access, i.e., the inbuilt functionalities set the limits to how users can navigate and analyse historical materials. They provide the tools and heuristics via which content becomes visible and users can create their research corpus or data set (see below).

However, while such type of access generally works for humanities’ scholars like historians, it does not necessarily meet the needs of those who follow more data-driven approaches, such as computational humanities researchers, computational historians, or NLP researchers. The latter often wish to process larger datasets for automatic enrichment and filtering, among other tasks. While most libraries or platforms provide access via search, accessing “newspaper collections as data” (i.e., at scale) is becoming more prevalent, but contemporary portals remain limited in their support for such interactions with the data.

**Toxicity and cultural bias.** As newspapers are embedded in specific spatial and temporal contexts, their content also contains traces of historical biases, both in text and image. In the most extreme cases, historical newspapers contain “toxic” content, to use a term common in today’s research on language models and ethical AI. The textual (or visual depiction) of people, especially the more marginalised and underprivileged, articulate attitudes which are considered offensive within contemporary norms.

But not all biased language is “toxic”. A more neutral term would be “overrepresentation” of specific textual patterns among certain subsets of the data, for example conservative newspapers may mention words such as “agriculture” more frequently than newspapers of other political leaning.

#### ► Measures to help achieve transparency and fairness

Related to the activity of “research corpus creation, selection and sharing”, there are a number of measures that could help or improve transparency and fairness. These measures are both at the level of the cultural heritage institutions providing the digitised newspaper collections as well as at the level of the researchers who create their research corpora.

**Collection documentation.** Providers of digitised historical newspapers, such as cultural heritage institutions and specific newspaper portals (e.g., ANNO, BelgicaPress, Chronicling America, Delpher, *impresso*, NewsEye, etc.) can provide detailed information regarding the collection. For example: list of newspaper titles, dates of publication, how much of a newspaper title has been digitised. It could also be useful to provide whatever contextual information about a newspaper and the entire collection is available, e.g., concerning selection criteria, geographical scope, number of editions, print runs, publishers and editors. Information such as political orientation of the newspaper titles, even when imperfect and tied to specific time periods, will be useful here. Ideally, such contextual information is accompanied by sources such as bibliographic references. The question of who is responsible for providing this information was raised, e.g., the cultural heritage institution or the researcher undertaking the research. Perhaps a partnership between these two actors would be most valuable.

Figure 18 illustrates how contextual information could be displayed: firstly, the Newspaper Timelines<sup>55</sup> from the *impresso* project, and secondly, the Press Picker<sup>56</sup> from the Living with Machines project.

The provision of explicit information regarding digitisation quality (e.g., Optical Character Recognition, OCR) would also be useful, ideally provided at a number of levels: for the whole newspaper title, for an issue, or per article.

**Terms of use.** Digitised newspaper providers should provide explicit guidelines regarding terms of use, particularly in terms of legal consideration. For example, the *impresso* platform requires users to sign a Non-Disclosure Agreement (NDA)<sup>57</sup> before access to full collection is granted. Furthermore, it would be useful for cultural heritage providers to provide information about what percentage of the total collection has been digitised. This helps to provide transparency on the “missingness” in a collection, ideally at the level of each of the newspaper titles.

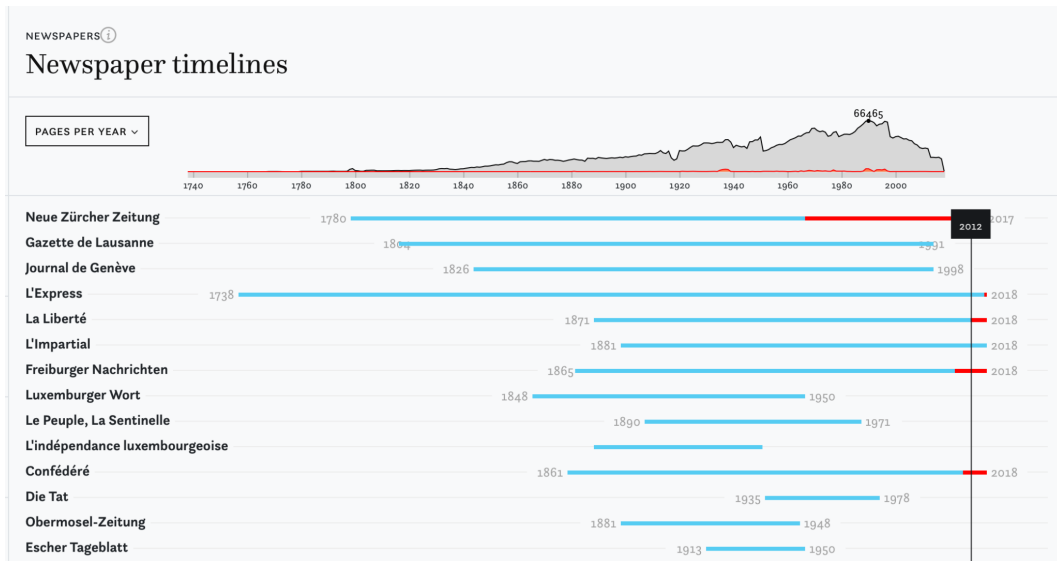
**Contested terms.** When considering measures to ensure transparency and fairness of research corpus creation, selection and sharing, it is important to consider diversity, equality, equity. To assist both researchers and cultural heritage institutions with this, an equity monitor could be developed. A number of aspects could be considered; for example, the identification of contested terms in a corpus (see [11, 6], as well as [3] and Conconcor<sup>58</sup>). If collection holders are aware that their corpora include contested terms, a disclaimer

<sup>55</sup> <https://impresso-project.ch/app/newspapers/>

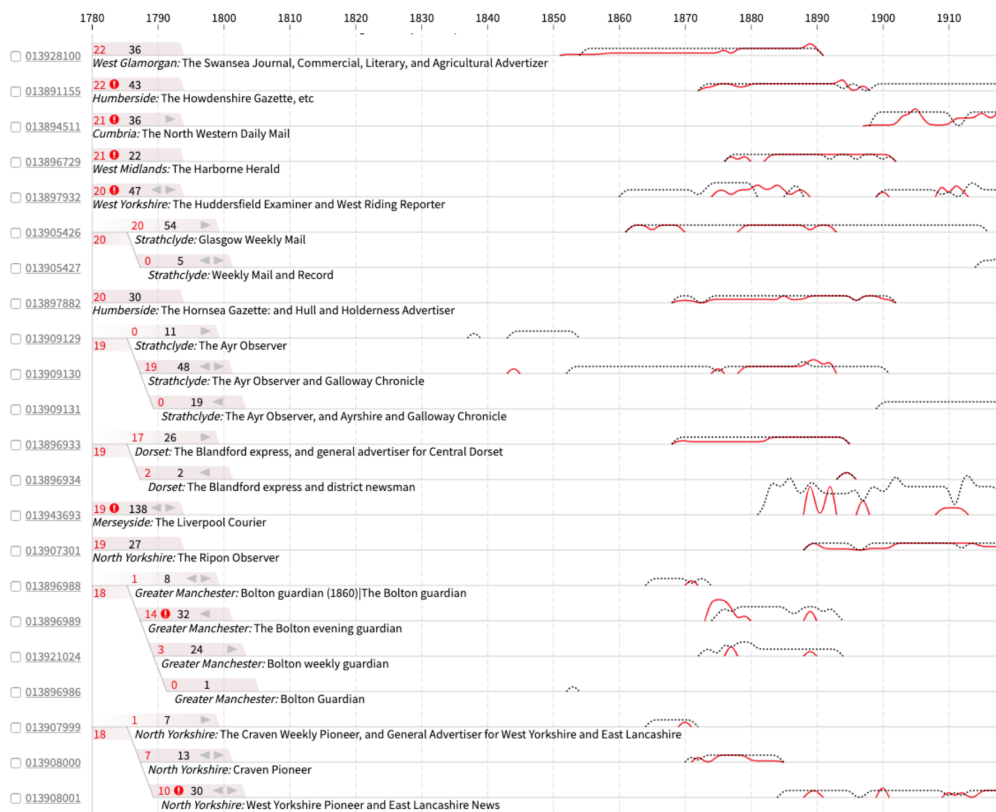
<sup>56</sup> <https://livingwithmachines.ac.uk/press-picker-visualising-formats-and-title-name-changes-in-the-british-librarys-newspaper-holdings/>

<sup>57</sup> [https://impresso-project.ch/assets/documents/impresso\\_NDA.pdf](https://impresso-project.ch/assets/documents/impresso_NDA.pdf)

<sup>58</sup> <https://www.cultural-ai.nl/conconcor>



(a) Timeline overview of available newspaper on the *impresso* platform (© impresso).



(b) Timeline overview of British newspapers on the PressPicker tool (© Living With Machines).

■ **Figure 18** Top: Timeline overview of available newspaper on the *impresso* platform (© impresso); Bottom: Timeline overview of British newspapers on the PressPicker tool (© Living With Machines).

alerting users to this could be added to the website. This could be particularly relevant when contentious terms are used for query expansion or are used in a visualisation. For researchers whose corpus includes contested terms, it is advisable to explicitly acknowledge this in their publications.

### 4.3.7 Processing and enrichment

#### ► Focus areas for transparency and fairness

In almost all cases, a raw, digitised newspaper corpus is processed in several ways before it is suitable for analysis. This could include, for example, Optical Character Recognition (OCR), Optical Layout Recognition (OLR), lexical processing such as part-of-speech tagging, named entity recognition (NER), linking to knowledge graphs, topic detection, or sentiment analysis. It could also include manual annotation of, for example, topics, people, or viewpoints. Each processing and enrichment step introduces bias or unwanted over- and under-representations into the data. When data is retrieved via a search engine or recommendation system, the ranking algorithm of that system also plays a role.

**Tool performance.** Regarding “low” level processing tasks (OCR/OLR), bias is mainly related to the quality of the tools. Do they work equally well on each part of the collection? How does OCR/OLR quality impact the retrievability and accessibility of each (type of) document? The quality of the tools on each part of the collection will depend on the data that they were originally trained on or developed for. They will likely work best on data that resembles the training/development set.

Similarly, for higher level, automated enrichment tasks (part-of-speech tagging, NER, linking to knowledge graphs, topic detection, sentiment analysis), we can ask: how well do they work for each part of the collection? What data were the tools trained on or developed for, and to what extent is this different from the data currently under investigation? NER tools may show a higher performance on some entities than others. Knowledge graphs may not cover all relevant entities.

**Ingrained bias.** In some cases, bias is ingrained in the collection [3]. As a product of their times, historical newspapers will also reflect the norms, values and language of e.g. colonising nations and their perspectives on their colonies. The use of automated enrichment tools may lead to unwanted side effects with respect to colonial or otherwise outdated terminology. Words may be taken out of context. Consider, for example, that a topic detection algorithm may define a geographically-focused topic as a list of terms including racist references to people.

**Posterior annotation.** Manual annotation is prone to bias that relates to the viewpoints, background and knowledge of the annotator. In some cases, these highly personal characteristics will be unknown, such as when making use of crowdsourcing. In some cases, we might not want to expose anonymous crowd workers to bias that is ingrained in the collection, such as when offensive, colonial terminology is present.

**Ranking.** Search engines typically rank documents based on a combination of the following factors: a matching score between a query and the content of the document, usage data in the form of previous queries and clicks, and an importance score of the document, e.g., using a PageRank-like algorithm. Whether a document will appear high in the ranking will therefore be influenced by its popularity (if usage data is taken into account), connectedness (if PageRank is used), and document properties such as length, which impact the matching score [20]. This may introduce distorted perceptions of search results. Therefore ranking-algorithms should be made transparent to users who rely on them.

► **Measures to help achieve transparency and fairness**

**Fine-grained OCR performance metrics.** Detailed information about OCR and segmentation quality helps a user to decide not only whether the quality is good enough, but also whether bias towards certain parts of the research corpus is to be expected. This requires fine-grained performance metrics, for example at the level of articles, newspaper titles or time periods.

**Access to “raw” data.** Another solution to mitigate or at least understand bias due to OCR errors is to provide access to the original “raw” scans, i.e., the images of pages.

**Systematic documentation of tools and training sets.** For automated enrichment tools, documentation of how, for what purpose, and on which training set they were created, helps a user to assess whether bias is to be expected when these tools are applied to their data. Several documentation approaches have been proposed in recent years, the most notable being Datasheets for Datasets [6] for documentation of training sets, and Model Cards [10] for documentation of trained models. Also, the research on provenance is relevant here, i.e. a formal representation of the consecutive processes involved in the creation of the enrichments. PROV<sup>59</sup> is an approach to formally capture provenance information on the semantic web.

**Scanning for contentious terms.** The content of historic newspaper collections will often be “biased” in the sense that the articles display the perspectives of the time in which they were created. Removing this type of bias will mostly not be feasible or desirable in the context of historical research. We recommend to include an “equity monitor” as part of a research design, where a user critically assesses whether contentious terminology is present in the corpus, and whether this is problematic. As noted above, contentious terminology could be problematic when used as input to automated enrichment tools, or when presented to crowd workers. In these cases, a user could decide to not include certain articles in their research. Note that detection of contentious terminology is not trivial and automation of this task is still in its infancy [3].

**Disclaimer about contentious language.** A user may include a disclaimer as part of the dataset, to warn (other) users and/or annotators that there may be offensive content. This is especially recommended when sharing the corpus for reproducibility or future research.

**Representative annotators.** We consider human annotators to be always biased. A diverse or representative group of annotators helps to avoid annotations that are skewed towards one background or viewpoint.

**Transparency about annotators.** Explicit information about who created the annotations helps users to assess whether an (unwanted) bias in the annotations is to be expected. This could consist of age, gender, country of citizenship, (native) language, level of expertise, and way of recruitment of the annotator(s). Note that this information is often not available when using crowdsourcing.

**Multiple relevance rankings.** Bias introduced by the ranking algorithm of a search engine may be explicated and mitigated by providing multiple rankings. Many search engines already include additional rankings next to relevance ranking, such as a chronological order. However, specifically the inclusion of multiple relevance rankings would allow a user to understand to what extent the ranking algorithm impacts their goals.

---

<sup>59</sup> <https://www.w3.org/2001/sw/wiki/PROV>

### 4.3.8 Data analysis

Once a research corpus has been selected and processed it is ready for analysis. Analysis already is an integral part of data selection, processing and enrichment. At first, we discussed this stage as part of the processing and enrichment stage. There are, however, tasks that clearly come after data processing and enrichment; for instance, the identification of named entities in the corpus has to be undertaken with NER software. Therefore, we have decided to identify it as a separate step in the research workflow.

Depending on the research question and the skill set of researchers, the analysis may be performed on the entire collection (e.g., downloading all the data and analysing it with a Jupyter notebook) or of a subset of the collection generated via the search and filter options in a portal. It also may be performed qualitatively, with a scholar browsing and reading specific articles, or quantitatively, applying computational approaches and tools.

#### ► Focus areas for transparency and fairness

**Traceability.** In order to make the research traceable, researchers have to be explicit about the methods and tools they use including, for the latter, the version, the used settings, and why they were appropriate for the task in question.

Often, the analysis of a newspaper corpus involves a set of tools organised in a pipeline. In order to obtain transparency, users should have insight into the composition and performance of the various components of the pipeline. In the case of machine learning tools, it should be clear which versions are used and on which dataset they have been trained.

In order to be fully transparent, ideally all these things are documented, and the tools or queries stored alongside the data. This raises the question what level of documentation is required to make the research traceable or repeatable for others. Some researchers provide tools to document the settings, such as the Gephi Fieldnotes plugin developed by [22]. Others have proposed strategies for tracing all the data handling steps [7]. We see, however, the risk that the effort to produce such documentation may take a disproportionate amount of time and effort. Tool standardisation may make this need less urgent (e.g., the role of SPSS<sup>60</sup> in Social Science research) and therefore reduce this burden.

#### ► Measures to help achieve transparency and fairness

**Access to facsimiles.** To facilitate qualitative research, where users explore the corpus at object level, an interface should present research results in the form of scans next to the OCR and metadata (which most portals currently afford). For quantitative analysis, tools will be used both inside and outside the portals. In order to improve the traceability of the research, researchers should have the ability to store tools and their settings alongside the datasets, perhaps on publicly accessible platforms such as GitHub and Zenodo, or using tools specifically designed for this purpose, such as the Gephi Fieldnotes plugin. “How to cite” text blocks with detailed and multimodal information on the tools and their settings could be helpful here.

**Comparative perspectives.** The transparency of analysis pipelines can be supported by interfaces that allow researchers to compare the performances of different algorithms on a specific task. An example is the NEWSGAC platform<sup>61</sup> that allows users to compare different

---

<sup>60</sup><https://www.ibm.com/de-de/analytics/spss-statistics-software>

<sup>61</sup><https://github.com/newsgac/platform>

algorithms for automatic genre detection in newspapers. In order to increase the transparency of machine learning tools, references to publications on models used and documentation on training datasets is provided (e.g., in the form of “datasheets for datasets”[6]).

**Replication.** Ideally, users should have the possibility to save, export and reuse their own analysis pipeline and the results. This output should ideally connect to the changing publication and presentation modes for the research results, that allow researchers to include data, code and narratives alongside each other (e.g., the Journal of Digital History<sup>62</sup> that publishes the data alongside the narrative and a description of the methodological issues, or the ESWC conference<sup>63</sup> where linking to data, code and other resources is a review requirement); this is further discussed in the Acknowledgements section below.

### 4.3.9 Visualisation of data, results and bias thereof

Visualisations have the added value of showing complex matter e.g. in graphs and images that help users to get a better overview of collections, corpora, data sets and also the content of these. Visualisations are important because they support the exchange between data and users since they can help to contextualise the collection on the one hand and research on the other. Using graphs, timelines, charts, maps, word-clouds, bubbles, and similar transparency concerning the collection and the research method is offered. The possibilities of how to support transparency by visualising collections and data span a wide range: e.g., research questions and methods are made explicit, topics can be contextualised, a classification of genres and faults or missing/biased data in the collection can be made visible, OCR-/layout-quality and research approaches can be identified, and comparisons of topics (and many other similar things) are possible. As a consequence interfaces can be designed to offer possibilities for visualisation.

#### ► Focus areas for transparency and fairness

**General Guidelines for Visual Design.** Visual interfaces are in many contexts suitable, necessary means to make patterns inherent in the data set in question salient to the observer. However, visualisations are abstract representations of (typically) numerical data, and the visual mapping of the underlying information always imposes a level of distortion because numbers are rather easier to compare when they are served in textual form than when our brain has to approximate them when they appear in the form of visuals such as bars in a bar chart or dots in a scatter plot. The complexity of comprehending data in textual form increases with the size of the data set, but visualisations help to arrange the data in a way that users get a quick, understandable overview even for vast data sets. In order to limit the level of distortion, visual representations of data have to be carefully designed.

**Accurate representations of data** . Following Edward Tufte’s guidelines for graphical excellence, first of all, visualisations should “show the data”, make it coherent and avoid distorting what the data has to say [19]. An appropriate indicator for good visual design is when viewers are induced to think about the substance rather than about methodology, graphic design or the technology of graphic production. Moreover, visualisations should not “lie,” i.e., the size of effect shown in the visual display needs to correspond to the size of effect in the data.

<sup>62</sup> <https://journalofdigitalhistory.org/en/about>

<sup>63</sup> <https://2022.eswc-conferences.org/call-for-papers-research-track/>



**Choice of colour.** Of particular importance for visual design is the selection of appropriate colour maps. Qualitative colour maps (a set of different hues) should be used to display categorical data, and continuous colour maps (sequential or diverging) to communicate quantitative data (colour gradients). Although powerful, a general advice is not to encode the most important feature with colour (“get it right in black and white”). Visualisations should also be colorblind-safe, i.e., one should not mix diverse shades of green and red. Several online tools support defining accurate colour maps, e.g. ColorBrewer<sup>64</sup>.

**Clarity.** Next to choosing inappropriate colour maps, visualisation designers should avoid visual clutter that reduces the readability of the displayed data and conceals occurring patterns. Clutter can also occur when choosing 3D over 2D representations, which are the means of choice when visualising data that does not inhere 3D structures. Textures that cause visual stress (moiré vibrations) should furthermore be avoided.

**Visual Exploration.** In order to support Information seeking, visualisation tools should implement Shneiderman’s mantra “Overview first, zoom and filter, then details on demand” [17]. Whereas the overview corresponds in digital humanities terminology to distant reading, details on demand refers to close reading. Thus, visual interfaces should support gradual zooming and filtering of the data to be analysed.

**Visualising uncertainty.** Especially, data in the context of humanities applications often embody uncertainty of different kinds (imprecision, inhomogeneity, incompleteness). Visualisations are suitable means to communicate these uncertainties, for example through transparency or grey glyphs, indispensable for increased reliability of visual display of information.

#### ► Measures to help achieve transparency and fairness

**Participatory design.** To ensure a transparent visual interface with a minimised level of data distortion, we suggest conducting a participatory visual design process that involves visualisation experts on the one hand, and domain experts that ensure the suitability of the visual design for its intended purpose on the other. An exhaustive overview of visual design principles can be found in [12].

**Collection visualisation.** There are some good examples on how interfaces can offer transparency on the collections. The CLARIAH-NL Media Suite<sup>65</sup>, and the Sound and Vision Archive<sup>66</sup>, offer explanations and graphs to contextualise the audiovisual collection of the Netherlands Institute for Sound and Vision (NISV), explaining both the role of the archive in the archival field in the Netherlands as well as the time the collection spans, what the digital archives cover, what kind of media is included, when updates happen, what part or the collection is digitised, how the collection is enriched, where additional information and help can be found and how it can be searched (as exemplified in Figure 19). It also indicates what the differences between the metadata of two media management systems are and offers links as well as downloads to the description of the metadata fields.

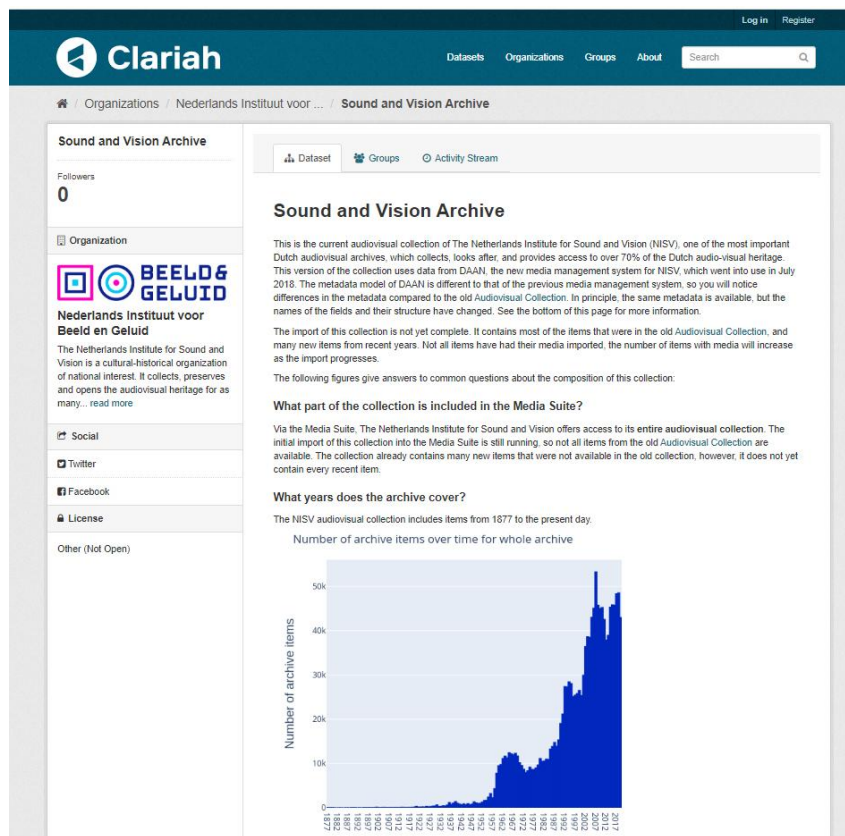
Digital newspaper collections also contextualise their datasets to a certain extent although additional information like the one offered for the Sound and Vision archive might be added. The *impresso* interface indicates the provenance of its collection in the detail view of

---

<sup>64</sup> <https://colorbrewer2.org/>

<sup>65</sup> <https://mediasuitedata.clariah.nl/dataset/nisv-catalogue>

<sup>66</sup> <https://mediasuitedata.clariah.nl/dataset/audiovisual-collection-daan>



■ **Figure 19** Landing Page of the Sound and Vision Archive which is one of the datasets of the Nederlands Instituut voor Beeld en Geluid within CLARIAH (© Nederlands Instituut voor Beeld en Geluid).

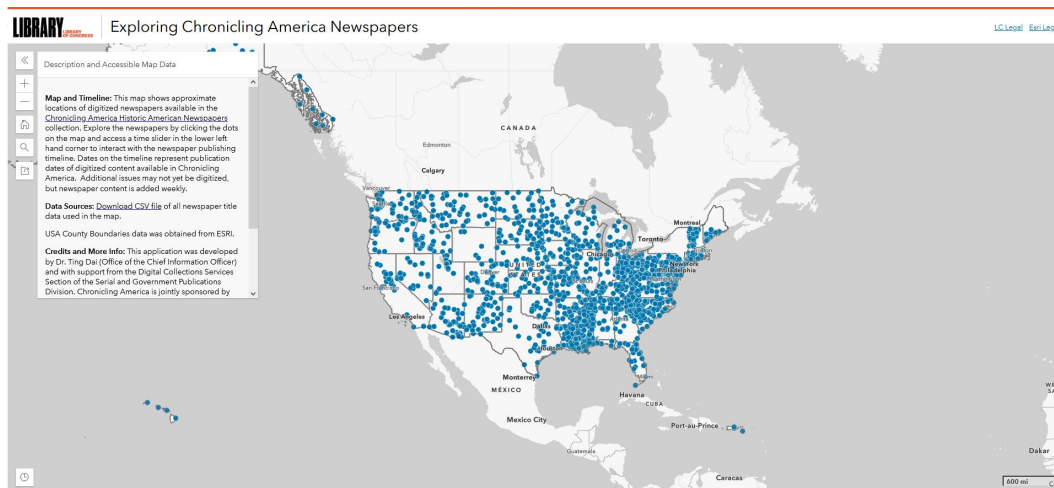
individual titles. A simple graph visualises the overall temporal distribution of the newspaper collection and a bar chart enables users to get a quick overview of the time span and amount of data each newspaper contributes to the collection (see Figure 17). Due to the efforts of the National Digital Newspaper Program<sup>67</sup> which “is a long-term effort to develop an Internet-based, searchable database of U.S. newspapers with descriptive information and select digitization of historic pages” Chronicling America is currently also adding information on its dataset. Figure 20 shows that contextualisation regarding the collection is made in the “Maps and Visualisation” section of the portal and it can be seen that most of this information is of very recent date (mostly updated February and June 2022) and that the “Map and Timeline” of the collection are updated on a weekly basis using the ArcGIS Instant App (see example in Figure 21). The interactive map visualisation has the added value that it supports a scalable reading of the collection, which is often required by humanities researchers.

These examples show that awareness for the necessity of transparency and biases is growing constantly. In this context visualisations can be helpful to support the communication of complex issues at a glance. The visualisation in Figure 16 frames the issue at hand very

<sup>67</sup> <https://www.loc.gov/ndnp/>

The screenshot shows the Library of Congress website interface. At the top, there are navigation buttons for 'ASK A LIBRARIAN', 'DIGITAL COLLECTIONS', and 'LIBRARY CATALOGS', along with a search bar containing 'Search Loc.gov' and a 'GO' button. Below this is a breadcrumb trail: 'The Library of Congress > National Digital Newspaper Program > Chronicling America Maps and Visualizations'. The main content area is titled 'Chronicling America Maps and Visualizations' and includes a 'Print', 'Subscribe', 'Share/Save', and 'Give Feedback' menu. A sidebar on the left contains a search box and a list of links: 'NDNP Home', 'About the Program', 'Guidelines & Resources', 'Award Recipients', 'Program News', 'NDNP Extras', and 'Contact the NDNP'. Below the sidebar, there is a section for 'Chronicling America' with a brief description and links to 'Go to Chronicling America', 'Topics in Chronicling America', and 'Chronicling America Maps and Visualizations'. The main content area lists several interactive maps and visualizations, such as 'Exploring Chronicling America Newspapers: All Digitized Titles (Map and Timeline) Updated Weekly', 'Chronicling America Temporal Coverage - Entire Collection (Area Chart) Updated July 2022', and 'Chronicling America Temporal Coverage by State (Map) Updated July 2022'. At the bottom, there is a section for 'Exploring Chronicling America Newspapers: All Digitized Titles (Map and Timeline) Updated Weekly' with a brief description of the visualization.

■ **Figure 20** Screenshot of “Chronicling America Maps and Visualizations” where information about the collection can be found (© Library of Congress).



■ **Figure 21** Screenshot of the visualisation of Chronicling America Newspapers, an interactive map created using the ArcGIS Instant App which enables an in depth exploration of the newspapers available in the collection ranked by the primary place of publication (© Library of Congress).

clearly. It is a good example of how well visualisations are able to communicate difficult and complex information. This is also true for analyses that are being done by researchers (see Section Data analysis).

#### 4.3.10 Acknowledgement

##### ► Focus areas for transparency and fairness

**Reveal hidden labour in digitisation.** There is a considerable amount of hidden labour in the process of generating digital newspaper collections. It would be ideal if this hidden labour could be made visible and all the actors in the process could be acknowledged for their work. This would help increase the level of transparency and fairness of the whole digitised historical newspaper ecosystem.

In order for historical newspaper collections to exist, they have to have been acquired by libraries, archives or museums, often in physical form. They then need to be digitised, processed (e.g. metadata generation, creation of lower resolution images for public display), and then enriched (e.g., OCR and article segmentation). It is not feasible to undertake the whole digitisation workflow at once, and therefore selection or prioritisation is needed. All these steps involve the expertise of cultural heritage professionals, computer and data scientists and software engineers, as well as (digital) humanities researchers who would like to use the digitised newspapers as historical sources for their research.

**Reveal hidden labour in data curation and enrichment.** In addition, the creation of research data sets or corpora, which requires sustained intellectual effort, is often not recognised or acknowledged as formal research output. This both discourages researchers to spend time and properly document this crucial step in the research workflow, but also devalues this work as a necessary evil before the “real” research work can begin. Both cultural heritage professionals and computer scientists contribute significantly to this phase, e.g., by providing historical context information about the historical newspaper collections or by working on information extraction methods to computationally facilitate the corpus building process. It is therefore important that this valuable work is uncovered and made visible.

Finally, such work is often made possible by the financial contribution of (public) funding agencies, the acknowledgement of which is also important.

##### ► Measures to help achieve transparency and fairness

**Acknowledgement in portals.** We recommend the formal and visible acknowledgement of all contributing partners at each step of the digitisation process. This includes the contributions by cultural heritage professionals, software engineers, researchers, and (public) funding agencies. For example, when building a platform for the exploration and analysis of digitised historical newspapers, both the funding agency can be acknowledged, such as is the case with the *impresso* project (“*impresso*. Media Monitoring of the Past. Supported by the Swiss National Science Foundation under grant CR- SII5 173719, 2019.”<sup>68</sup>) or the *NewsEye* project (“This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 770299.”<sup>69</sup>). Additionally, when building a research dataset or corpus, it is important to acknowledge all contributors, such as “Biltreyst, Daniël, Philippe Meers, Dries Moreels, Julia Noordegraaf and Christophe Verbruggen. *Cinema Belgica: Database for Belgian Film History*.”<sup>70</sup>. Not only does this publicly acknowledge the work of people involved in the development of the platform or

<sup>68</sup> <https://impresso-project.ch>

<sup>69</sup> <https://www.newseye.eu/about/>

<sup>70</sup> <https://www.cinemabelgica.be>, all consulted on July 27, 2022

dataset in question, but it also enables it to be cited in articles or other research outputs that have made use of it. Being able to demonstrate the impact of such platforms becomes increasingly important for their sustainability.

**Acknowledgement in publications.** Acknowledgement in publications is another key method to make all parties who contributed to the development of data and the design of platforms visible. For instance, the domains in which each actor has contributed to such a dataset or platform could be stated explicitly, as is already required with regard to authorship by some journals. The Journal of Open Humanities Data Author Guidelines<sup>71</sup> for example provides a number of recommendations based on the ICMJE (Internal Committee of Medical Journal Editors)<sup>72</sup>, outlining criteria for authorship. An area where there is still room for further development is the (academic) recognition of more innovative digital research outputs. Innovative Journals such as the Journal of Open Humanities Data (JOHD)<sup>73</sup>, which focuses on the publication of “peer reviewed publications describing humanities data or techniques with high potential for reuse”<sup>74</sup>; the Journal of Digital History which intends “serve as a forum for critical debate and discussion in the field of digital history by offering an innovative publication platform and promoting a new form of data-driven scholarship and of transmedia storytelling in the historical sciences”<sup>75</sup> and the Journal of Data Mining & Digital Humanities (JDMDH)<sup>76</sup> which is situated at “the intersection of computing and the disciplines of the humanities, with tools provided by computing such as data visualisation, information retrieval, statistics, text mining by publishing scholarly work beyond the traditional humanities.”<sup>77</sup>

#### 4.3.11 Training

Training is essential to raise awareness of the complexity of enriched historical sources and variability in the quality of available data. In addition, and as we have outlined above, historical newspaper data may reproduce biases present in past societies. Training can provide the necessary understanding and tools to help deal with this.

**Digital literacy.** Training already exists on various levels. More and more, digital literacy is a part of the curriculum of (digital) humanities students. Some newspaper portals also offer domain-specific training. For example, *impresso* offers 1) General training on research using digitised historical newspapers, 2) Training on how to use the *impresso* processing tools and 3) Platform specific training about the functionality of the interface, linked to the FAQ, which contains references to literature.<sup>78</sup> *NewsEye* provides access to various training materials for schools and universities as well as material targeted at a general audience.<sup>79</sup> The *Ranke 2* platform, for example, offers a series of lessons on Digital Source criticism that can be helpful not only for students.<sup>80</sup>

<sup>71</sup> <https://openhumanitiesdata.metajnl.com/about/submissions/>

<sup>72</sup> <https://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html>

<sup>73</sup> <https://openhumanitiesdata.metajnl.com>

<sup>74</sup> <https://openhumanitiesdata.metajnl.com/about/>

<sup>75</sup> <https://www.degruyter.com/journal/key/jdh/html>

<sup>76</sup> <https://jdmdh.episciences.org>

<sup>77</sup> <https://jdmdh.episciences.org/page/editorial-policies>

<sup>78</sup> <https://impresso-project.ch/theapp/usage/>

<sup>79</sup> <https://www.newseye.eu/>

<sup>80</sup> <https://ranke2.uni.lu/>

■ **Table 5** Stages of digital humanities workflows for historical newspapers including focus areas and measures to implement.

	Focus area	Measures
<b>Research corpus</b>	Awareness of dig. & pres. policies Diverse user needs Toxicity and cultural bias	Collection documentation Multiple data access points Terms and conditions Contested terms
<b>Processing &amp; Enrichment</b>	Tool performance  Ingrained bias  Posterior annotation  Ranking	Performance metrics Access to “raw” data Doc. of tools and training sets Scanning for contentious terms Disclaimers Representative annotators Transparency about annotators Offer multiple relevance rankings
<b>Data analysis</b>	Traceability	Access to facsimiles Replication Comparative perspectives
<b>Visualisation</b>	Visual design guidelines Accurate data representations Colour choices Clarity Visual exploration Uncertainty	Collection visualisation Participatory design
<b>Acknowledgement</b>	Reveal hidden labour in dig., data curation and enrichment	Ackn. in portals and publications
<b>Training</b>	Digital literacy	Publications with best practices Code examples Example workflows and use cases Platform-specific training API training

We identified the following types of training as contributing to the skills and knowledge of researchers with respect to transparency and fairness when studying digitised historical newspapers:

1. Publications with best practices;
2. Code examples, for example in the form of Jupyter notebooks;
3. Example workflows and example use cases;
4. Example lesson plans and course material;
5. Platform specific training, in the form of interface walk-throughs;
6. API documentation.

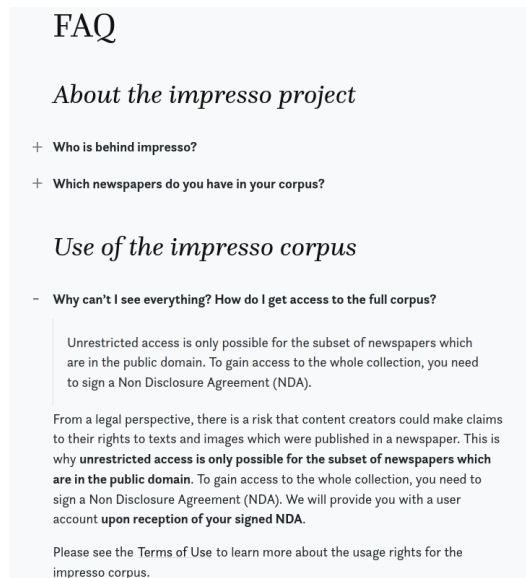
Table 5 offers a high-level overview of the proposed focus areas and measures to achieve transparency and fairness.

In the following last section we undertake a review of the *impresso* portal and assess to which extent it fulfils the above-mentioned criteria for transparency and fairness.



### 4.3.12 Application: Analysis of *impresso* portal

We revisited the vaccination case study and its underlying questions to apply it to the *impresso* interface for historical newspapers which was developed in close cooperation between historians, computer scientists and designers and with special emphasis on transparency. We revisited the focus areas and accompanying measures we identified above and concentrated on opportunities for improvement of the current interface.



■ **Figure 22** FAQ page on the *impresso* website (© *impresso*).

*Corpus creation and selection.* The Newspapers overview page offers a good overview of the print runs of the newspapers in the collection. Information about missing pages and issues (mismatches between available data and expected data based on library metadata) is available but could be explained more clearly. Information on the origins of the collections from different partners is available albeit scattered across the interface and could be further improved by links to the respective digitisation policies of the partnering institutions. *impresso* offers rich metadata for individual titles (e.g. name, print run, number of pages, orientation, regional focus) and links to their respective pages on the websites of the institutions from which they originate. The value of this information could be increased by allowing users to use them as filters in the search component and by offering a download of the data for individual processing and analysis outside the interface. More detailed information about the alignment between *impresso*'s collection and the collection of the partnering institutions should be added. The hitherto unsolved problem to relate the “tip of the iceberg” of the available digital content to the rest of it, i.e., the total record of newspapers in circulation in the past persists in the *impresso* interface as well. Search results are sorted by “relevance” by default, the underlying settings are not explained. Access is granted via a browser interface and following approval after signing an Non-Disclosure-Agreement. Users are able to export metadata for articles and, depending on legal agreements, also full text. The corresponding FAQ entry should be expanded to explain the content of the export file.

*Processing, enrichment, and analysis of data.* The FAQ section of the interface collects important information about the project, corpus and interface. Semantic enrichments such as topic modelling, named entity recognition or text reuse are well explained as is the entry



on OCR quality. Filtering by OCR quality should be enabled. The FAQ entries for data export and legal restrictions are valuable but could more concisely explain the legal status of exported data and link directly to project publications on system architecture and the processing pipeline.

*Visualisation of results.* The application makes use of data visualisations in multiple forms which are accompanied by corresponding FAQ entries. This includes frequently distributions over time e.g. for search results, graphs to represent overlaps between topics. The Inspect&Compare<sup>81</sup> component uses small multiples of bar charts to reveal overlaps and dissimilarities between two queries or article collections and can e.g. be used to evaluate search strategies [1]. The interface is clearly designed for research purposes. Basic knowledge about data and interactive visualisations is a prerequisite.

*Training and digital literacy.* From a user perspective, the portal supports both scholars with low digital literacy (as tools for analysis are built in) and for more advanced skilled researchers (because data and metadata can be exported as csv). The project has compiled and integrated educational materials for researchers ranging from beginner to advanced level which cover digitised newspapers per se as well as the functionalities of the interface.<sup>82</sup> Tutorials could be placed more prominently in the interface and specific support for visually impaired users is missing.

*Acknowledgements.* As stated above, the project indicates the source of funding by the Swiss National Science Foundation together with an overview of the full project consortium<sup>83</sup> and the contributions of individuals.

#### 4.3.13 Conclusions

It was our goal to compose a set of recommendations for content providers such as libraries and archives as well as developers of research interfaces, in order to help individual researchers in the field to gain as much transparency and fairness as is required for the analysis of digitised and enriched historical newspaper collections. We did so by focusing on aspects with a potentially high impact on the outcome of research. We found that efforts to increase transparency and fairness have to be made on all stages of the workflow. The stages we identified were 1) corpus creation and selection; 2) processing, enrichment, and analysis of data; 3) visualisation of results; 4) training and digital literacy; and 5) acknowledgements. These stages interrelate and are dependent on each other. It was therefore not always possible to make clear distinctions. We discussed issues of transparency and fairness in each of the stages and also reflected on some measures that can help mitigate biases, lacks of transparency and fairness.

Overall it can be summarised that we found a lot of variability in the current landscape of digital newspapers. This applies within and across newspaper collections and includes differences in metadata standards such as METS/ALTO, the quality of OCR with older processing tending towards lower quality, but also in regard to the scope of enrichment: whereas some content providers invested in the correct identification of even small news items (e.g., obituaries), others only offer PDFs of scanned images. These variations have a high impact on research but are still poorly communicated in contemporary interfaces.

For researchers, therefore, some challenges remain and some of these challenges can be countered by content and interface providers. For example, it remains crucial to understand the “digital sample”, i.e., researchers have to be given the ability to assess the (non-)

<sup>81</sup> <https://impresso-project.ch/app/inspect>

<sup>82</sup> <https://impresso-project.ch/theapp/usage/>

<sup>83</sup> <https://impresso-project.ch/consortium/people/>

representativity of the small number of digitised and available newspapers against the background of all past and potentially not archived newspapers. Also, providing detailed information – such as metadata, information on the digitisation process, distribution over time, political orientation of newspapers, links to historical contextual information, etc. – of the nature of the collections that can be found via the interfaces, can be of great help to researchers. Interfaces should therefore provide intuitive guidelines for and explanation of the collection. Also, information regarding diversity, equality, and equity (such as a notification of contested terminology) should be found.

Since the quality of OCR and layout recognition as well as classification issues remain a challenge, it also remains crucial for researchers to be able to always have access to the “raw” data, i.e., the images. If automated tools or training sets are available, a proper documentation has to be provided. The experimental nature of the analysis tools still poses challenges regarding replication and sustainability (e.g., the query storage facility, notebooks for re-training on the same corpora). For example, the CLARIAH Media Suite allows users to rerun queries but updates to the underlying collections cause this feature to break. Again, the progress in automation creates a need for extensive documentation (e.g., the Gephi Fieldnotes Plugin) which could be diminished once there is some standardisation of the tools (e.g., as in the case of SPSS that is more integrated in a methodological framework and also generally accepted as reliable). Another option could be to allow researchers to compare the results of different algorithms. For many research questions, however, it also remains important to work with external tools and adaptable methods. The possibility to download data sets and analyse on (or publish about) them outside the interfaces is one more important feature we identified. Overall, a higher degree of transparency in visual interfaces can be a real asset for humanities research.

## References

- 1 Düring, M., Kalyakin, R., Bunout, E. & Guido, D. Impresso Inspect and Compare. Visual Comparison of Semantically Enriched Historical Newspaper Articles. *Information*. **12**, 348 (2021,9), <https://www.mdpi.com/2078-2489/12/9/348>, Number: 9 Publisher: Multidisciplinary Digital Publishing Institute
- 2 Beelen, K., Lawrence, J., Wilson, D. & Beavan, D. Bias and representativeness in digitized newspaper collections: Introducing the environmental scan. *Digital Scholarship In The Humanities*. (2022)
- 3 Brate, R., Nesterov, A., Vogelmann, V., Van Ossenbruggen, J., Hollink, L. & Van Erp, M. Capturing Contentiousness: Constructing the Contentious Terms in Context Corpus. *Proceedings Of The 11th On Knowledge Capture Conference*. pp. 17-24 (2021)
- 4 Ehrmann, M., Bunout, E. & Düring, M. Historical Newspaper User Interfaces: A Review. *Proceedings Of The 85th International Federation Of Library Associations And Institutions (IFLA) General Conference And Assembly*. pp. 24 (2019), <https://infoscience.epfl.ch/record/270246?ln=en>
- 5 Fry, B. Visualizing data. (" O'Reilly Media, Inc.",2008)
- 6 Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J., Wallach, H., Iii, H. & Crawford, K. Datasheets for datasets. *Communications Of The ACM*. **64**, 86-92 (2021)
- 7 Hoekstra, R. & Koolen, M. Data scopes for digital history research. *Historical Methods: A Journal Of Quantitative And Interdisciplinary History*. **52**, 79-94 (2019)
- 8 Linhares Pontes, E., Cabrera-Diego, L., Moreno, J., Boros, E., Hamdi, A., Doucet, A., Sidere, N. & Coustaty, M. MELHISSA: a multilingual entity linking architecture for historical press articles. *International Journal On Digital Libraries*. **23**, 133-160 (2022)
- 9 McGillivray, B., Poibeau, T. & Ruiz Fabo, P. Digital humanities and natural language processing “Je t’aime ... Moi non plus”. (Alliance of Digital Humanities,2020)

- 10 Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. & Gebru, T. Model cards for model reporting. *Proceedings Of The Conference On Fairness, Accountability, And Transparency*. pp. 220-229 (2019)
- 11 Modest, W. & Lelijveld, R. Words Matter, Works in Progress I. National Museum of World Cultures. (2018)
- 12 Munzner, T. Visualization analysis and design. (CRC press,2014)
- 13 Neudecker, C., Baierer, K., Gerber, M., Clausner, C., Antonacopoulos, A. & Pletschacher, S. A survey of OCR evaluation tools and metrics. *The 6th International Workshop On Historical Document Imaging And Processing*. pp. 13-18 (2021)
- 14 Oberbichler, S., Boroş, E., Doucet, A., Marjanen, J., Pfanzelter, E., Rautiainen, J., Toivonen, H. & Tolonen, M. Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians. *Journal Of The Association For Information Science And Technology*. **73**, 225-239 (2022)
- 15 Hechl, S., Langlais, P., Marjanen, J., Oberbichler, S. and Pfanzelter, E., Digital interfaces of historical newspapers: opportunities, restrictions and recommendations. *Journal Of Data Mining & Digital Humanities*. (2021)
- 16 Schneider, P. Rerunning OCR: A Machine Learning Approach to Quality Assessment and Enhancement Prediction. *ArXiv Preprint ArXiv:2110.01661*. (2021)
- 17 Shneiderman, B. The eyes have it: A task by data type taxonomy for information visualizations. *The Craft Of Information Visualization*. pp. 364-371 (2003)
- 18 Van Strien, D., Beelen, K., Ardanuy, M., Hosseini, K., McGillivray, B. & Colavizza, G. Assessing the impact of OCR quality on downstream NLP tasks. (SCITEPRESS-Science,2020)
- 19 Tufte Edward, R. The visual display of quantitative information. (Cheshire, Connecticut: Graphic Press,2001)
- 20 Traub, M., Samar, T., Van Ossenbruggen, J., He, J., Vries, A. and Hardman, L. Querylog-based assessment of retrievability bias in a large newspaper corpus. *2016 IEEE/ACM Joint Conference On Digital Libraries (JCDL)*. pp. 7-16 (2016)
- 21 Vrijenhoek, S., Kaya, M., Metoui, N., Möller, J., Odijk, D. and Helberger, N. Recommenders with a mission: assessing diversity in news recommendations. *Proceedings Of The 2021 Conference On Human Information Interaction And Retrieval*. pp. 173-183 (2021)
- 22 Wieringa, M., Geenen, D., Es, K. and Nuss, J. The Fieldnotes Plugin: Making Network Visualization in Gephi accountable. *Good Data*. **14** pp. 277 (1988)

#### 4.4 Towards an International Historical Newspaper Infrastructure

*Clemens Neudecker (Staatsbibliothek zu Berlin, DE)*

*Maud Ehrmann (EPFL – Lausanne, CH)*

*Matteo Romanello (EPFL – Lausanne, CH)*

*Martin Volk (Universität Zürich, CH)*

*Lars Wieneke (C2DH – Esch-sur-Alzette, LU)*

*Dario Kampkaspar (TU Darmstadt, DE)*

**License** © Creative Commons BY 4.0 International license

© Clemens Neudecker, Maud Ehrmann, Dario Kampkaspar, Matteo Romanello, Martin Volk, and Lars Wieneke

Portals and platforms that aggregate digitised newspapers from multiple sources and institutions have added great value for researchers, as they e.g. allow the comparative study of newspaper data from different countries and in multiple languages from within a uniform

environment. With few exceptions, digitised newspapers are commonly made available online via national portals and collections, or even fragmented across numerous online repositories, with each offering different features and functionalities for searching and accessing the data. Together, this makes working with digitised newspapers very tedious for researchers, and raises the need for standardised and modular information flows between systems [6]. Additionally, new tools and services are required for the accommodation of scholarly research requirements for content discovery and management, and to reflect their iterative, exploratory research workflows.

Different types of actors act and collaborate in the field of digitised newspapers. Libraries are predominantly interested in the continued digitisation, preservation and online presentation of their newspaper holdings, often providing only very basic ways to browse the digitised newspapers by title or date, and to a lesser extent, perform keyword searches when full text has been produced using Optical Character Recognition (OCR). Researchers, on the other hand, are in need of more advanced ways to access the data, build corpora from it, or explore it through quantitative aspects. Computer scientists, research software engineers and designers, finally, work on developing robust text and image processing approaches and scholarship-oriented (re)search interfaces. All actors contribute to advancing many aspects of historical newspaper access, processing and research, some focusing on breadth (e.g., libraries taking care of full collections), others on depth (e.g., research projects venturing into innovative prototypes on small or medium-sized collections).

Research interests are as diverse as the newspaper's contents, with use cases from different disciplines such as the humanities, social sciences, computational linguistics, computer science, or digital humanities. Similarly, a multitude of methods are currently used in the analysis and exploration of digitised newspapers either as images, text or a combination thereof. Furthermore, specific approaches are typically required for document and language processing to deal with challenges due to the primarily historical and multilingual nature of the newspaper content.

In an ideal world, portals like Europeana Newspapers<sup>84</sup> and *impresso*<sup>85</sup> would continue to aggregate additional newspaper data, and add more specialised functionalities for computational approaches to digitised historical newspapers.

Looking at the current situation, due to resource constraints, libraries struggle to offer more specialised functionalities as would be desired by researchers. The main focus of cultural heritage institutions involved with newspaper digitisation remains on increasing the volume of content available digitally for the general public, and providing long-term access to these digitised resources. Digital infrastructures that give access to digitised newspapers from multiple countries and in multiple languages are faced with the challenge of sustaining continued development, aggregation of additional newspaper content, and the integration of specific functionalities and interfaces for computational analysis of digitised newspapers, once funding runs out.

What could be simple and lightweight alternatives to better support a diverse research community interested in the computational analysis of digitised newspapers? And how could the cost and effort to start new research projects and collaborations around digitised newspapers be reduced? What is the essential required infrastructure, what improvements are achievable with reasonable time and effort and what could be a more ambitious, long-term vision for the international newspaper digitisation and research ecosystem?

---

<sup>84</sup> <https://www.europeana.eu/en/collections/topic/18-newspapers>

<sup>85</sup> <https://impresso-project.ch/app/>

Several initiatives have looked into cost-efficient ways to improve the provision of digitised newspapers for computational analysis. Already in 2013, Tim Sherratt coined the phrase “From portals to platforms” in his presentation at the LIANZA conference in New Zealand [1], suggesting that libraries should focus more on ways to publish their digitised data in machine-readable formats and via an Application Programming Interface (API), rather than building portals with sophisticated user interfaces. A main argument is that user interfaces will hardly ever be able to fully cater to all the diverse use cases, especially from researchers, as they are always constrained – pre-determined by a set of design decisions about what invites further exploration, or what is deemed necessary, relevant and useful, also for the greater public. Platforms that put a focus on ways for distributing data put the decisions back into the hands of the users, enabling them to obtain, interact with and analyse the data in their own preferred environment, and with the tools and methods of their own choice.

Another important perspective here is that of the US project *Collections as Data* [2]. The Principal Investigator of *Collections as Data* project, Thomas Padilla, was invited to participate in this Dagstuhl Seminar, but unfortunately could not attend. But the main ideas and concepts of *Collections as Data* were nevertheless present throughout the whole seminar, and especially in Sally Chambers evening lecture “Newspapers as Data: Challenges and Solutions”. A core ambition of *Collections as Data* lies in the use of practical and cost-efficient means to better support computational research of cultural heritage data. This also includes the call to provide “actionable” collections, via e.g. Jupyter notebooks and the use of APIs [4], or making the data available in ways that allow using it for machine learning purposes [3].

An intermediate (ideal?) solution that retains the best of each approach would be to develop and maintain interfaces that offer both capabilities, with innovative, powerful, and appropriate functionalities for content search, discovery, and comparison on the one hand, and on the other (or on the back-end of the former) user-facing data dumps and APIs.

To summarise, even without cross-national newspaper portals offering advanced search and exploration tools for researchers (or prior to their development), stakeholders, that is to say libraries, computer scientists and digital humanists, can still contribute to improving digitised newspaper collections for computational analysis by researchers in multiple, simple and practical, and sustainable ways. The below recommendations aim to provide some guidance on best practices for further improving the possibilities for computational approaches to digitised newspapers based on simple and practical means:

1. libraries should strive to always expose digitised newspaper content (metadata, images and full text) via API, preferably using the APIs from the International Image Interoperability Framework (IIIF)<sup>86</sup>; additionally, modalities to download bulk data (dumps) with different selections (only images, only OCR, OCR as plain text, by newspaper title or language etc) in simple, machine-readable formats (CSV, JSON, TXT) are seen as very useful by researchers across multiple disciplines. As these datasets are often substantial in size, delivery mechanisms – as established for the exchange of data in biology and physics – should be implemented;
2. the choice of (meta)data formats in libraries should follow de-facto standards in newspaper digitisation like the Library of Congress maintained XML-based standards METS/ALTO<sup>87</sup> or IIIF manifests;

---

<sup>86</sup> <https://iiif.io/>

<sup>87</sup> <http://www.loc.gov/standards/mets/>; <https://www.loc.gov/standards/alto/>

3. information on the provenance (selection, digitisation and processing) of digitised newspaper collections should be documented and made publicly available (e.g. via the *Atlas of Digitised Newspapers*<sup>88</sup>) to identify and assess biases or to find contextual information that is required to understand the background of how the data was produced and its implications on the use of it for scholarly research;
4. when new newspaper digitisation projects are conceived, they should always include layout analysis and OCR; as a next step, standardised ways for encoding information about image content in digital newspapers could be investigated;
5. the full text of digitised newspapers should be annotated with language labels based on automatic language identification. This is particularly important in multilingual countries like Belgium, Luxembourg, Italy, or Switzerland, as it has a potential impact on downstream natural language processing (NLP) tasks as well as on how contents are made searchable via user-facing exploration interfaces;
6. whenever possible, the full text of digitised newspapers should be enriched with named entity recognition (NER) and entity disambiguation and linking (EL) to a multilingual knowledge base such as e.g. Wikidata;
7. for all automatic data processing and enrichments like OCR, NER, EL etc., there should be information made available with the newspaper content that allows users to quickly assess the performance quality of this automatic processing to aid in selection of newspaper content for dataset and corpus building;
8. notation and interchange formats for tool processes and semantic enrichments should be standardised with the aim to achieve interoperability between tools, text and annotations (this could building on existing initiatives, e.g. the Distributed Text Services<sup>89</sup> and the know-how of libraries and historical newspaper research projects);
9. article segmentation, the detection of reading order and layout evaluation are difficult open questions that will require more work and research, and the development of metrics and tools for quality control which find community agreement;
10. a cross-national meta-search engine, catalogue or registry could help researchers find out more easily what newspapers are already digitised and with what granularity (e.g. scans only vs. OCR full text vs. article separation). Such additional metadata and contextual information could be added into established and sustainable newspaper catalogues and repositories (e.g. ZDB<sup>90</sup>);
11. to support the widest possible range of use and reuse scenarios, open licences such as the Public Domain Mark or Creative-Commons-Zero should be advertised and used for digitised newspapers, as even CC-BY-SA-NC or CC-BY-SA can be prohibitive for some relevant usages (e.g. the training and distribution of machine learning models [5]).

Step by step, and in collaboration with all stakeholders, the implementation of the above recommendations could form the basis for a “Newspaper Network Notation Framework” (N>NNF), similar to and following the same process as IIF, from basic design decisions in the wider community to concrete specifications, and eventually standardised implementations of interoperable newspaper collections that are fit for computational approaches.

---

<sup>88</sup> <https://www.digitisednewspapers.net/>

<sup>89</sup> <https://distributed-text-services.github.io/specifications/>

<sup>90</sup> <https://zdb-katalog.de/index.xhtml>

**References**

- 1 Sherratt, T. From portals to platforms. Building new frameworks for user engagement (2013)
- 2 Padilla, T. Responsible Operations: Data Science, Machine Learning, and AI in Libraries. *OCLC Research Position Paper* (2019)
- 3 Lee, B.C.G. The “Collections as ML Data” Checklist for Machine Learning & Cultural Heritage (2022)
- 4 Candela, G., Saez, M.-D., Escobar, P., Marco-Such, M. Reusing digital collections from GLAM institutions. *Journal of Information Science*, 48, 251-267 (2022, 2)
- 5 Lassner, D., Neudecker, C., Coburger, J., Baillot, A. Publishing an OCR ground truth data set for reuse in an unclear copyright setting. *Zeitschrift für digitale Geisteswissenschaften* (2021)
- 6 Romanello, M., Ehrmann, M., Clematide, S. & Guido, D. The Impresso System Architecture in a Nutshell. *EuropeanaTech Insights*. (2020), <https://pro.europeana.eu/page/issue-16-newspapers#the-impresso-system-architecture-in-a-shell>, <https://infoscience.epfl.ch/record/283595>



## Participants

- Kaspar Beelen  
The Alan Turing Institute –  
London, GB
- Estelle Bunout  
University of Luxembourg, LU
- Sally Chambers  
Ghent University, BE & KBR,  
Royal Library of Belgium –  
Brussels, BE
- Simon Clematide  
Universität Zürich, CH
- Mariona Coll-Ardanuy  
The Alan Turing Institute –  
London, GB
- Mickaël Coustaty  
University of La Rochelle, FR
- Marten Düring  
University of Luxembourg, LU
- Maud Ehrmann  
EPFL – Lausanne, CH
- Laura Hollink  
Centrum Wiskunde &  
Informatica – Amsterdam, NL
- Stefan Jänicke  
University of Southern Denmark –  
Odense, DK
- Axel Jean-Caurant  
University of La Rochelle, FR
- Dario Kampkaspar  
TU Darmstadt, DE
- Jana Keck  
German Historical Institute  
Washington, US
- Yves Maurer  
National Library of  
Luxembourg, LU
- Clemens Neudecker  
Staatsbibliothek zu Berlin, DE
- Julia Noordegraaf  
University of Amsterdam, NL
- Eva Pfanzelter  
Universität Innsbruck, AT
- David A. Smith  
Northeastern University –  
Boston, US
- Martin Volk  
Universität Zürich, CH
- Lars Wieneke  
C2DH – Esch-sur-Alzette, LU



## Remote Participants

- Antoine Doucet  
University of La Rochelle, FR
- Matteo Romanello  
University of Lausanne, CH